FINAL REPORT

Ultra-Low-Energy Sub-Threshold Circuits: Program Overview

PI: Anantha Chandrakasan

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

Final Report (ESE DARPA Program) - Summary:

Summary: In this DARPA program, we have developed a robust design methodology to scale power supply voltages to levels as low as 250mV, reducing the *energy dissipation of digital computation by an order of magnitude*. We have demonstrated both logic (standard cells) and memory. We have explored the use of parallelism to maintain performance at reduced power supply voltages. This concept was demonstrated with a UWB baseband processor. We have developed a DC-DC converter to efficiently deliver sub-threshold voltage and minimize the power dissipation of an arbitrary digital circuit. **We have demonstrated 9 test chips** in state-of-the-art 65nm, 90nm and 0.18µm CMOS technologies. All test chips were fabricated for free (primarily by TI).

The key outputs of this program are:

- Sub-threshold digital library was developed (for ASICs) in TI's 65-nm CMOS
  - □ 62 cells, Vdd ≈ 250 mV, (0 - 70°C), process variation tolerant
  - □ Integrated into commercial design flow and FIR filter test chip was demonstrated
- Developed Sub-VT 10T and 8T SRAM in 65-nm CMOS
  - □ Sense amplifier redundancy (8T) was employed to operate with supply of 350mV and data retention < 300mV
- Demonstrated a UWB radio baseband processor at Vdd = 400 mV
  - □ 100Mb/s @ 2.0 mW (20pJ/bit), 90-nm CMOS process
  - □ Parallelism employed (20X) : 25 MHz, 620 correlators, 4 matched filters
- Demonstrated a UWB Analog-to-Digital Converter
  - □ 500Msamples/sec, 65-nm CMOS process
  - □ Parallelism of 36 converters enables sub-threshold biasing
- Minimum energy tracking loop with embedded dc-dc converter
  - □ DC-DC converter delivers voltages down to 250 mV in 65nm CMOS
  - □ 50-100% energy savings by tracking & adjusting minimum energy point of operation
- Demonstrated a fully integrated switched sub-V_T DC-DC converter

Publications:

The following publications directly resulted from the DARPA ESE funding:

2006 ISSCC

1. B. Calhoun, A. Chandrakasan "A 256kb Sub-threshold SRAM in 65-nm CMOS"

2006 ICASSP

2. V. Sze, R. Blazquez, M. Bhardwaj, A. Chandrakasan, "An energy efficient sub threshold baseband processor architecture for pulsed ultra-wideband communications"

2006 IEEE Symposium on VLSI Circuits

3. B. Ginsburg, A. Chandrakasan, "A 500MS/s 5b ADC in 65nm CMOS"

2006 International Symposium on Low Power Electronics and Design

4. J. Kwong, A. Chandrakasan, "Variation-Driven Device Sizing for Minimum Energy Sub-threshold Circuits"
5. B. H. Calhoun, A. Wang, N. Verma, A. P. Chandrakasan, "Sub-threshold Design: The Challenges of Minimizing Circuit Energy" (invited)
6. B. P. Ginsburg and A. P. Chandrakasan, "500-MS/s 5-b ADC in 65-nm CMOS With Split Capacitor Array"2006 ISLPED Low Power Design Contest Award

2007 ISSCC

7. Y. Ramadass, A. Chandrakasan, "Minimum energy tracking loop with embedded dc-dc converter delivering voltages down to 250 mV in 65-nm CMOS" ISSCC 2007 Beatrice Winner Editorial Award
8. N. Verma, A. Chandrakasan, "A 65-nm sub-Vt SRAM employing sense-amplifier redundancy"
9. **Vivienne Sze, Anantha P. Chandrakasan,** "Design of an Ultra-Low Voltage UWB Baseband Processor" ISSCC/DAC Design Contest Award Winner

GOMACTech (Government Microcircuit Applications & Critical Technology Conference) 2007

10. B. Ginsburg, V. Sze, A.P. Chandrakasan, "A Parallel Energy Efficient 100Mbps Ultra-Wideband Radio Baseband," March 2007.
11. A. Wang, B. H. Calhoun, N. Verma, J. Kwong, A. Chandrakasan, "Ultra-Dynamic Voltage Scaling for Energy Starved Electronics," (poster presentation)

PESC 2007 (June 2007)

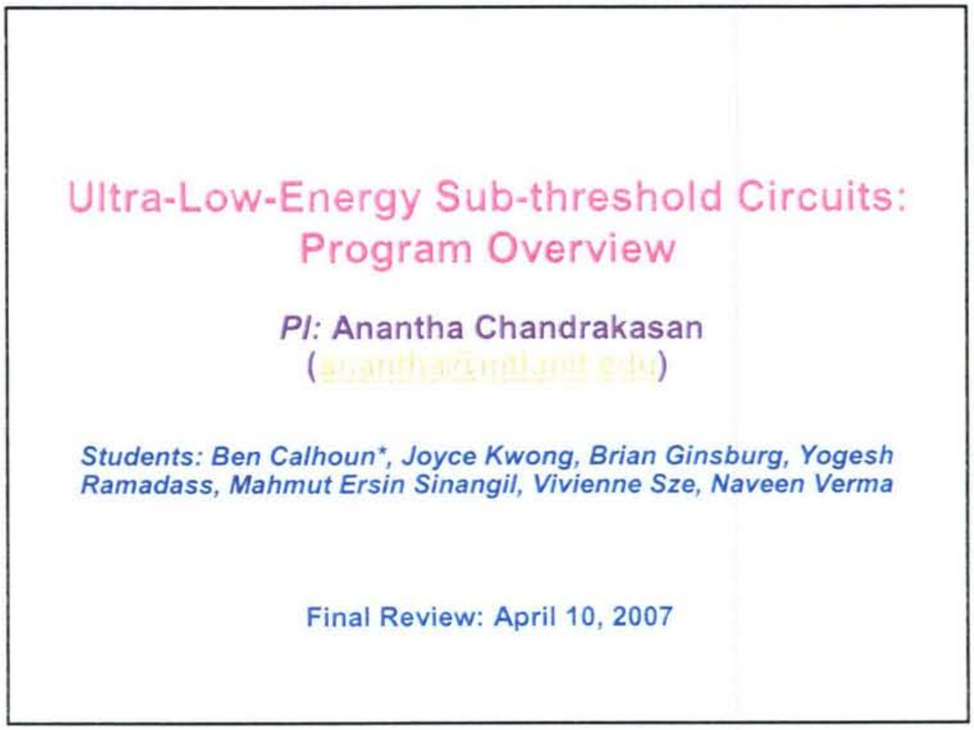
12. Yogesh K. Ramadass, Anantha P. Chandrakasan, "Voltage Scalable Switched Capacitor DC-DC Converter for Ultra-Low-Power On-chip Applications"

ISLPED 2007

13. V. Sze, A. Chandrakasan, "A 0.4V UWB Baseband Processor" to be presented

Journal Papers:

14. Benton H. Calhoun, Anantha P. Chandrakasan, "Static Noise Margin Variation for Sub-threshold SRAM in 65nm CMOS," IEEE Journal of Solid-State Circuits, vol. 41, no. 7, pp. 1673-1679, July 2006.
15. Benton H. Calhoun, Anantha P. Chandrakasan, "", IEEE Journal of Solid-State Circuits,
16. B. P. Ginsburg and A. P. Chandrakasan, "500-MS/s 5-bit ADC in 65-nm CMOS With Split Capacitor Array DAC," IEEE J. Solid-State Circuits, vol. 42, no. 4, pp. 739-747, Apr. 2007.
17. N. Verma, A. P. Chandrakasan, "A 256kb 65nm 8T Sub-Threshold SRAM employing Sense-Amplifier Redundancy", accepted to the IEEE J. Solid-State Circuits.

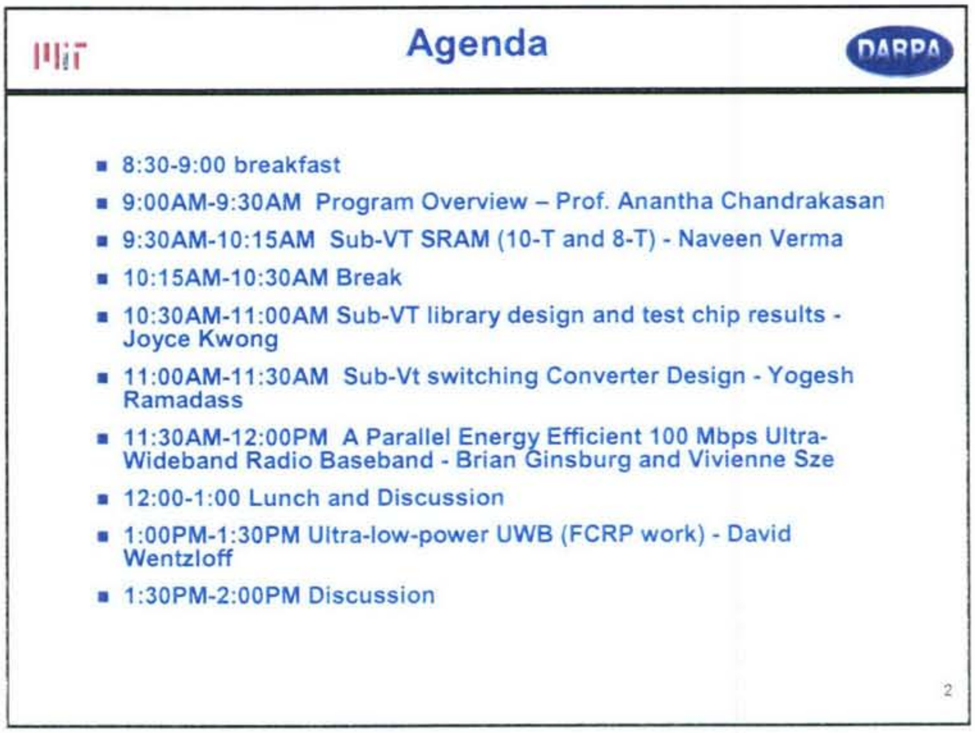


Ultra-Low-Energy Sub-threshold Circuits: Program Overview

PI: Anantha Chandrakasan
(anantha@mtl.mit.edu)

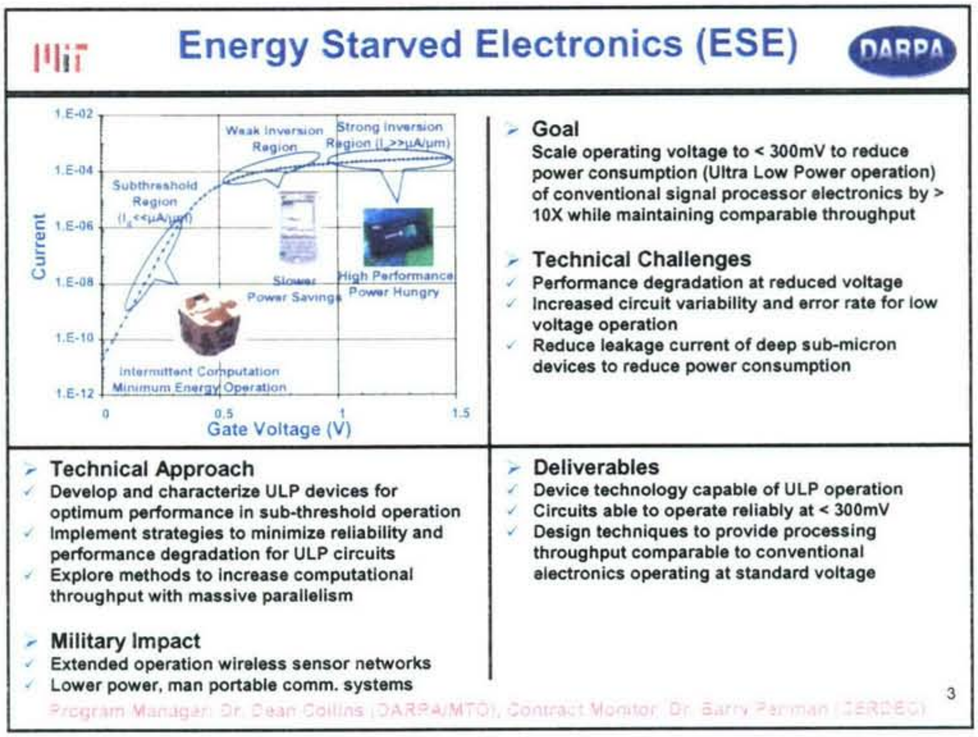
Students: Ben Calhoun, Joyce Kwong, Brian Ginsburg, Yogesh Ramadass, Mahmut Ersin Sinangil, Vivienne Sze, Naveen Verma*

Final Review: April 10, 2007



Agenda

- 8:30-9:00 breakfast
- 9:00AM-9:30AM Program Overview – Prof. Anantha Chandrakasan
- 9:30AM-10:15AM Sub-VT SRAM (10-T and 8-T) - Naveen Verma
- 10:15AM-10:30AM Break
- 10:30AM-11:00AM Sub-VT library design and test chip results - Joyce Kwong
- 11:00AM-11:30AM Sub-Vt switching Converter Design - Yogesh Ramadass
- 11:30AM-12:00PM A Parallel Energy Efficient 100 Mbps Ultra-Wideband Radio Baseband - Brian Ginsburg and Vivienne Sze
- 12:00-1:00 Lunch and Discussion
- 1:00PM-1:30PM Ultra-low-power UWB (FCRP work) - David Wentzloff
- 1:30PM-2:00PM Discussion


1.E-02
1.E-04
1.E-06
1.E-08
1.E-10
1.E-12
Current
0
0.5
1
1.5
Gate Voltage (V)
Weak Inversion Region
Strong Inversion Region (I_d>>µA/µm)
Subthreshold Region (I_d<<µA/µm)
Slower Power Savings
High Performance Power Hungry
Intermittent Computation Minimum Energy Operation

Energy Starved Electronics (ESE)

- **Goal**

  Scale operating voltage to < 300mV to reduce power consumption (Ultra Low Power operation) of conventional signal processor electronics by > 10X while maintaining comparable throughput

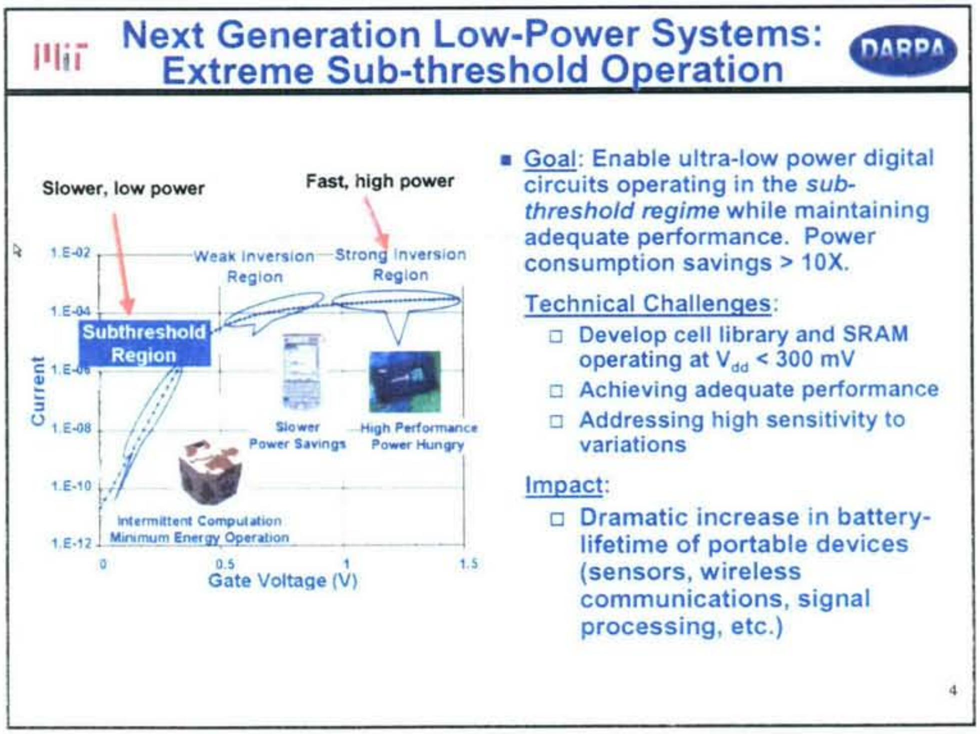
- **Technical Challenges**
  - ✓ Performance degradation at reduced voltage
  - ✓ Increased circuit variability and error rate for low voltage operation
  - ✓ Reduce leakage current of deep sub-micron devices to reduce power consumption

- **Technical Approach**
  - ✓ Develop and characterize ULP devices for optimum performance in sub-threshold operation
  - ✓ Implement strategies to minimize reliability and performance degradation for ULP circuits
  - ✓ Explore methods to increase computational throughput with massive parallelism

- **Military Impact**
  - ✓ Extended operation wireless sensor networks
  - ✓ Lower power, man portable comm. systems

- **Deliverables**
  - ✓ Device technology capable of ULP operation
  - ✓ Circuits able to operate reliably at < 300mV
  - ✓ Design techniques to provide processing throughput comparable to conventional electronics operating at standard voltage

Program Manager: Dr. Dean Collins (DARPA/MTO), Contract Monitor: Dr. Barry Perlman (CERDEC)

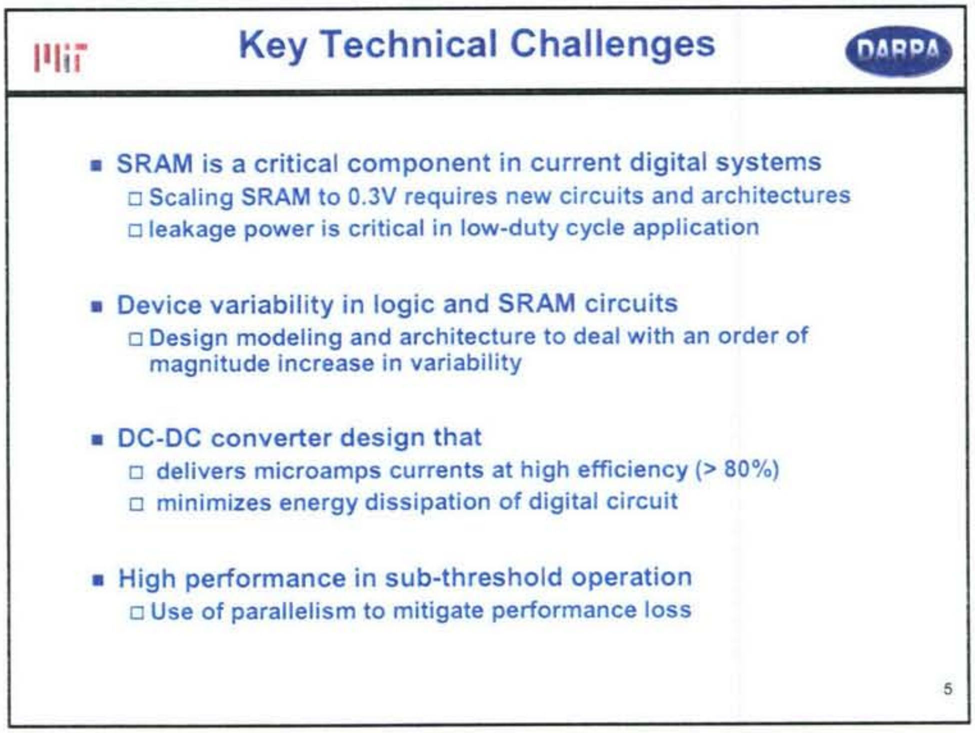

Slower, low power
Fast, high power
1.E-02
1.E-04
1.E-06
1.E-08
1.E-10
1.E-12
Current
0
0.5
1
1.5
Gate Voltage (V)
Weak Inversion Region
Strong Inversion Region
Subthreshold Region
Slower Power Savings
High Performance Power Hungry
Intermittent Computation Minimum Energy Operation

Next Generation Low-Power Systems: Extreme Sub-threshold Operation

- Goal: Enable ultra-low power digital circuits operating in the *sub-threshold regime* while maintaining adequate performance. Power consumption savings > 10X.

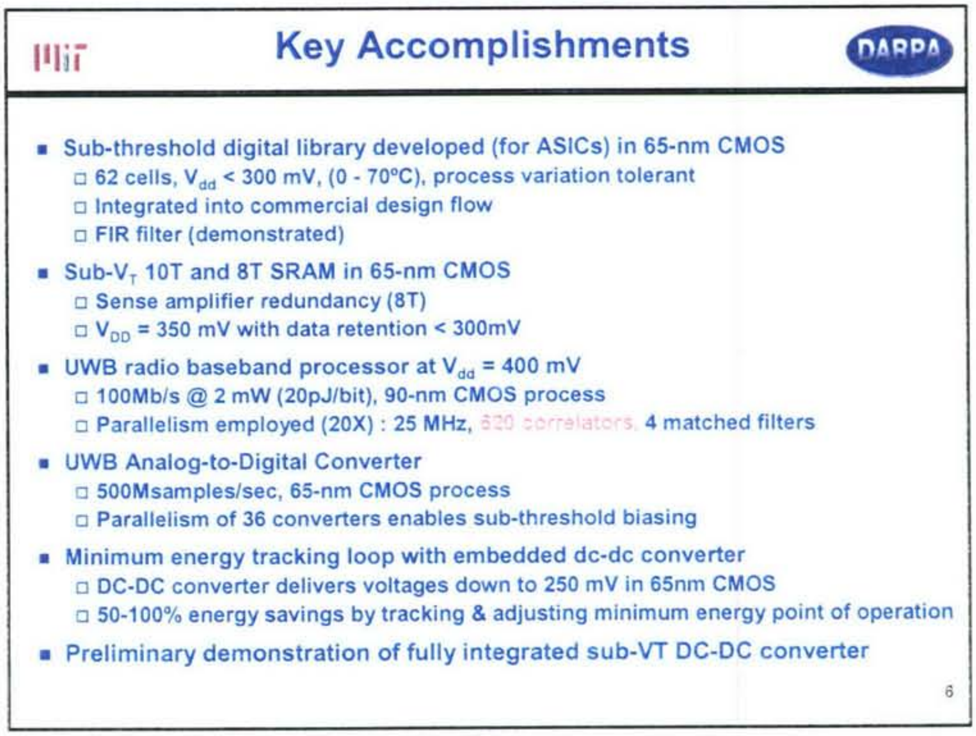
  Technical Challenges:
  - □ Develop cell library and SRAM operating at V_{dd} < 300 mV
  - □ Achieving adequate performance
  - □ Addressing high sensitivity to variations

  Impact:
  - □ Dramatic increase in battery-lifetime of portable devices (sensors, wireless communications, signal processing, etc.)



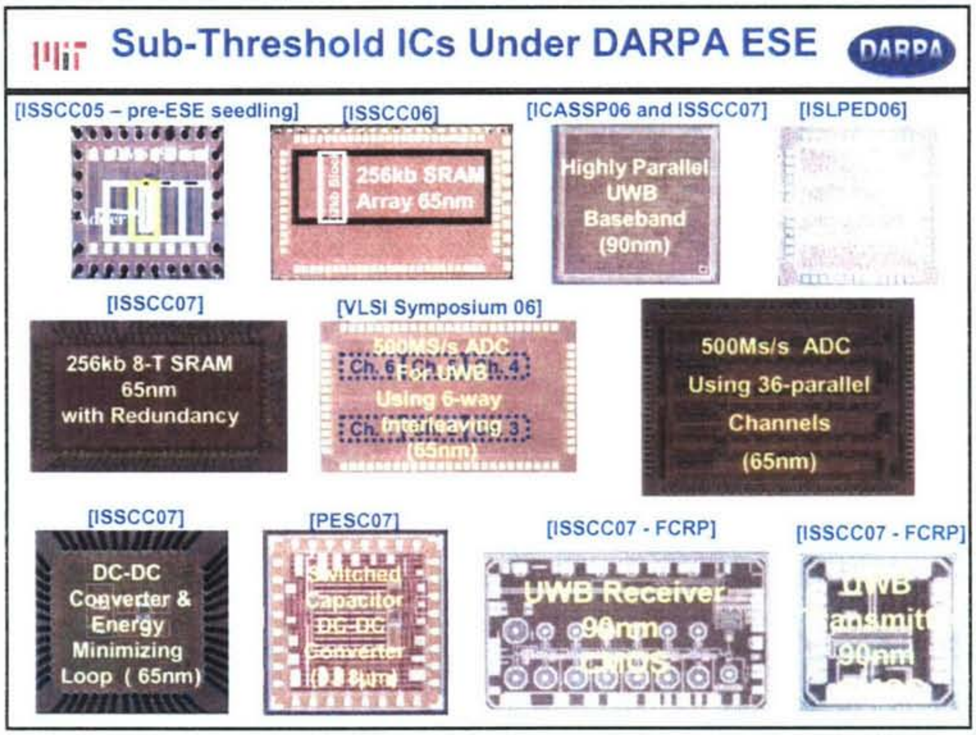
Key Technical Challenges

- SRAM is a critical component in current digital systems
  - □ Scaling SRAM to 0.3V requires new circuits and architectures
  - □ leakage power is critical in low-duty cycle application
- Device variability in logic and SRAM circuits
  - □ Design modeling and architecture to deal with an order of magnitude increase in variability
- DC-DC converter design that
  - □ delivers microamps currents at high efficiency (> 80%)
  - □ minimizes energy dissipation of digital circuit
- High performance in sub-threshold operation
  - □ Use of parallelism to mitigate performance loss



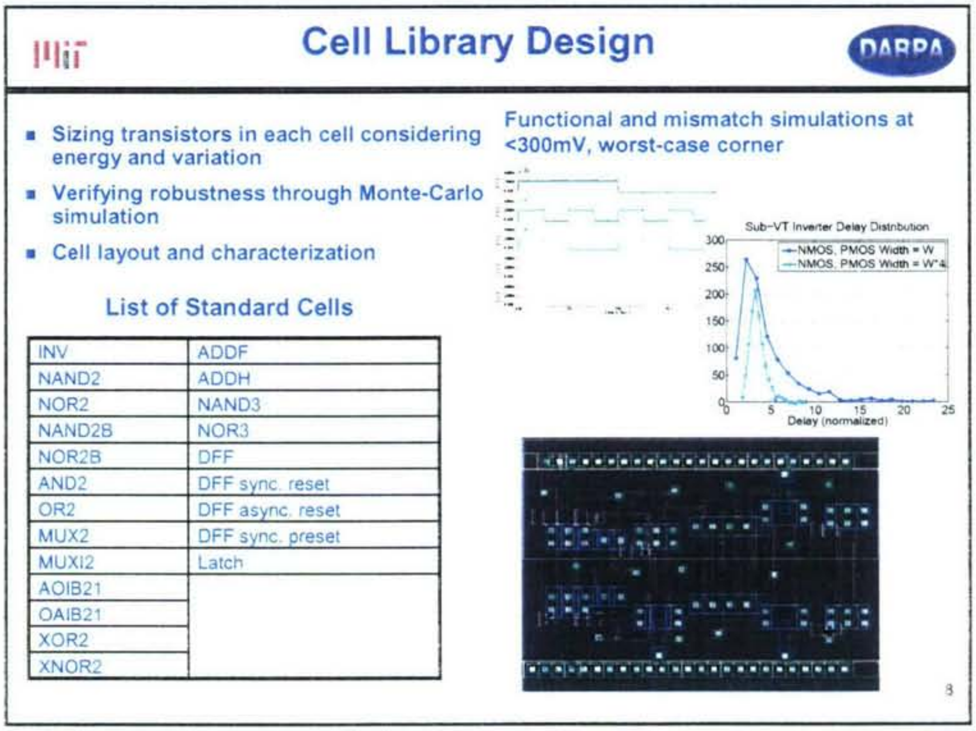
Key Accomplishments

- Sub-threshold digital library developed (for ASICs) in 65-nm CMOS
  - □ 62 cells, V_{dd} < 300 mV, (0 - 70ºC), process variation tolerant
  - □ Integrated into commercial design flow
  - □ FIR filter (demonstrated)
- Sub-V_T 10T and 8T SRAM in 65-nm CMOS
  - □ Sense amplifier redundancy (8T)
  - □ V_{DD} = 350 mV with data retention < 300mV
- UWB radio baseband processor at V_{dd} = 400 mV
  - □ 100Mb/s @ 2 mW (20pJ/bit), 90-nm CMOS process
  - □ Parallelism employed (20X) : 25 MHz, 620 correlators, 4 matched filters
- UWB Analog-to-Digital Converter
  - □ 500Msamples/sec, 65-nm CMOS process
  - □ Parallelism of 36 converters enables sub-threshold biasing
- Minimum energy tracking loop with embedded dc-dc converter
  - □ DC-DC converter delivers voltages down to 250 mV in 65nm CMOS
  - □ 50-100% energy savings by tracking & adjusting minimum energy point of operation
- Preliminary demonstration of fully integrated sub-VT DC-DC converter



Sub-Threshold ICs Under DARPA ESE

DARPA
[ISSCC05 – pre-ESE seedling]
Adder
[ISSCC06]
Sub Block
256kb SRAM Array 65nm
[ICASSP06 and ISSCC07]
Highly Parallel UWB Baseband (90nm)
[ISLPED06]
[ISSCC07]
256kb 8-T SRAM 65nm with Redundancy
[VLSI Symposium 06]
500MS/s ADC For UWB Using 6-way Interleaving (65nm)
Ch. 6 Ch. 5 Ch. 4
Ch. 3
500Ms/s ADC Using 36-parallel Channels (65nm)
[ISSCC07]
DC-DC Converter & Energy Minimizing Loop (65nm)
[PESC07]
Switched Capacitor DC-DC Converter (0.18μm)
[ISSCC07 - FCRP]
UWB Receiver 90nm CMOS
[ISSCC07 - FCRP]
UWB Transmitter 90nm



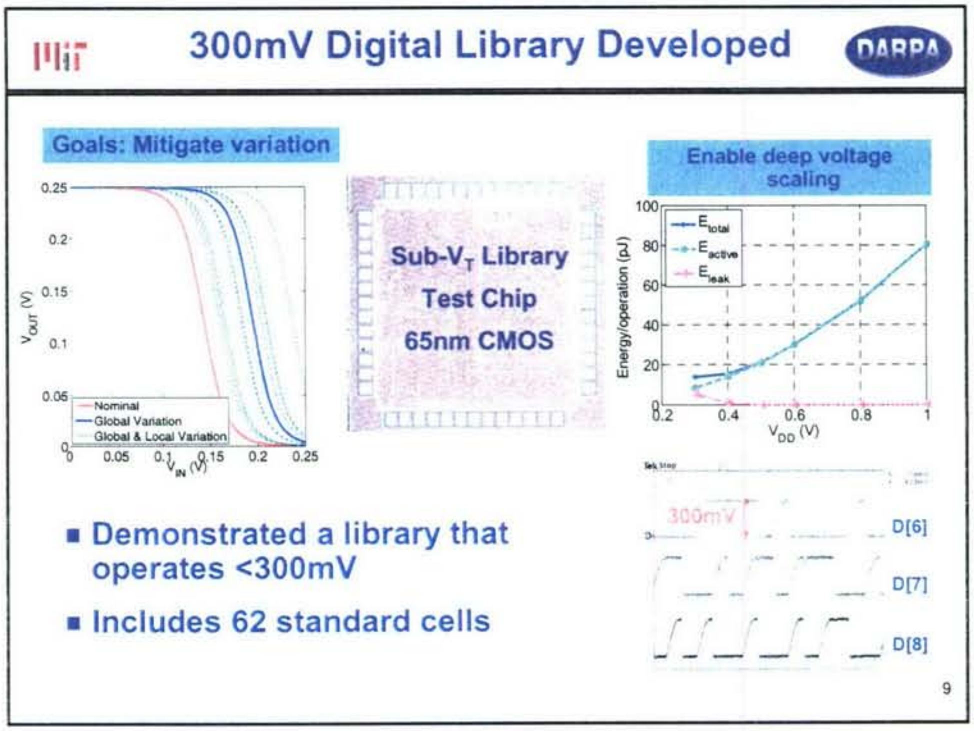
Cell Library Design

- Sizing transistors in each cell considering energy and variation
- Verifying robustness through Monte-Carlo simulation
- Cell layout and characterization

List of Standard Cells

INV	ADDF
NAND2	ADDH
NOR2	NAND3
NAND2B	NOR3
NOR2B	DFF
AND2	DFF sync. reset
OR2	DFF async. reset
MUX2	DFF sync. preset
MUXI2	Latch
AOIB21	
OAIB21	
XOR2	
XNOR2	

Functional and mismatch simulations at <300mV, worst-case corner



300mV Digital Library Developed

Goals: Mitigate variation

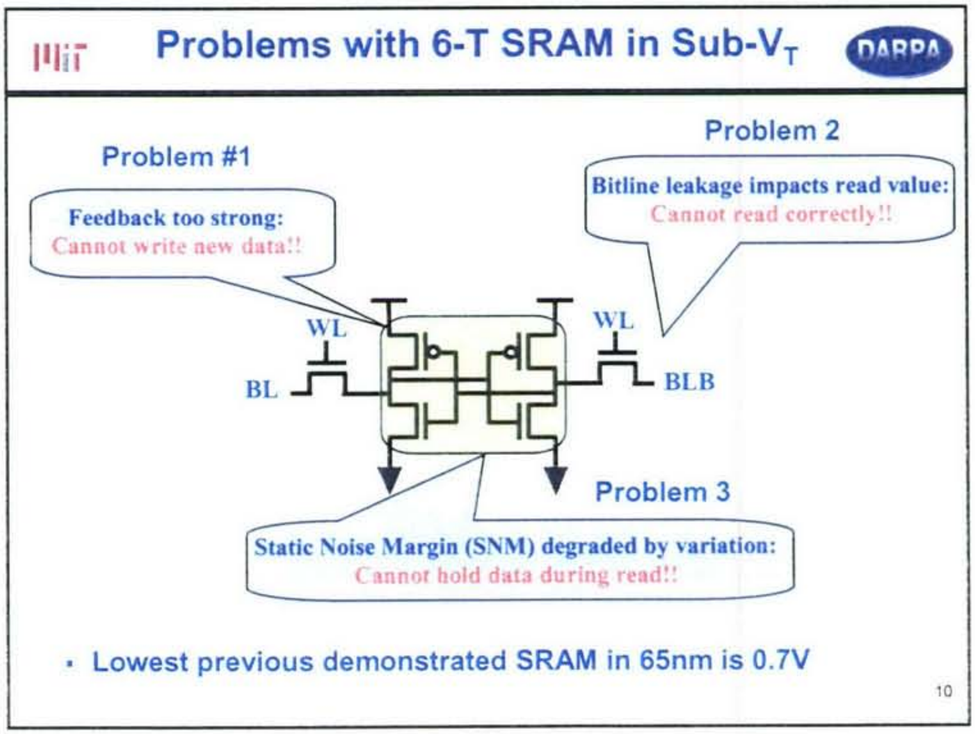
0.25
0.2
0.15
0.1
0.05
0
V_{OUT} (V)
0
0.05
0.1
0.15
0.2
0.25
V_{IN} (V)
Nominal
Global Variation
Global & Local Variation

Sub-V_T Library
Test Chip
65nm CMOS

Enable deep voltage scaling

100
80
60
40
20
0
Energy/operation (pJ)
0.2
0.4
0.6
0.8
1
V_{DD} (V)
E_{total}
E_{active}
E_{leak}
300mV
D[6]
D[7]
D[8]

- Demonstrated a library that operates <300mV
- Includes 62 standard cells



Problems with 6-T SRAM in Sub-V_T

Problem #1

Feedback too strong:
Cannot write new data!!

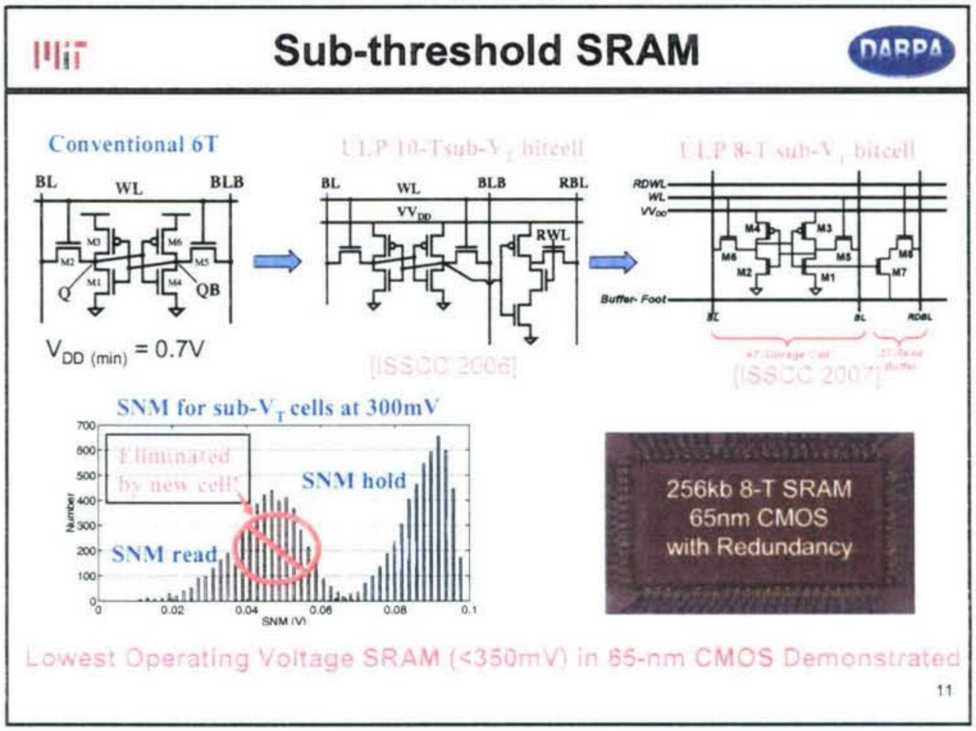
Problem 2

Bitline leakage impacts read value:
Cannot read correctly!!

Problem 3

Static Noise Margin (SNM) degraded by variation:
Cannot hold data during read!!

- Lowest previous demonstrated SRAM in 65nm is 0.7V



Sub-threshold SRAM

$V_{DD\ (min)}$ = 0.7V

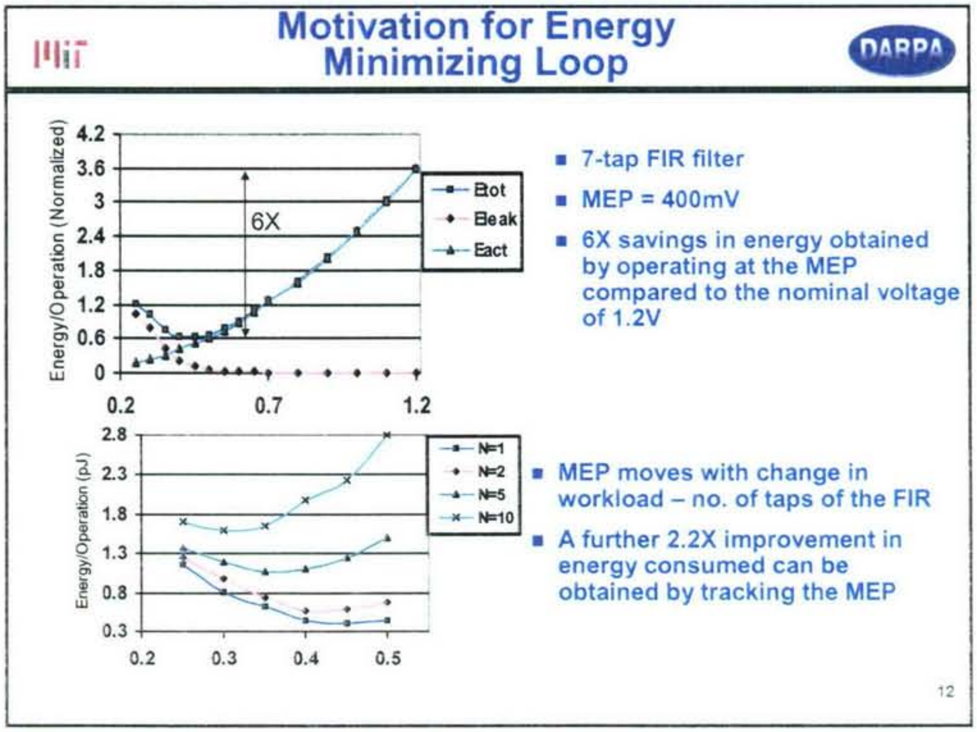
[ISSCC 2006]

[ISSCC 2007]

SNM for sub-V_T cells at 300mV

256kb 8-T SRAM
65nm CMOS
with Redundancy

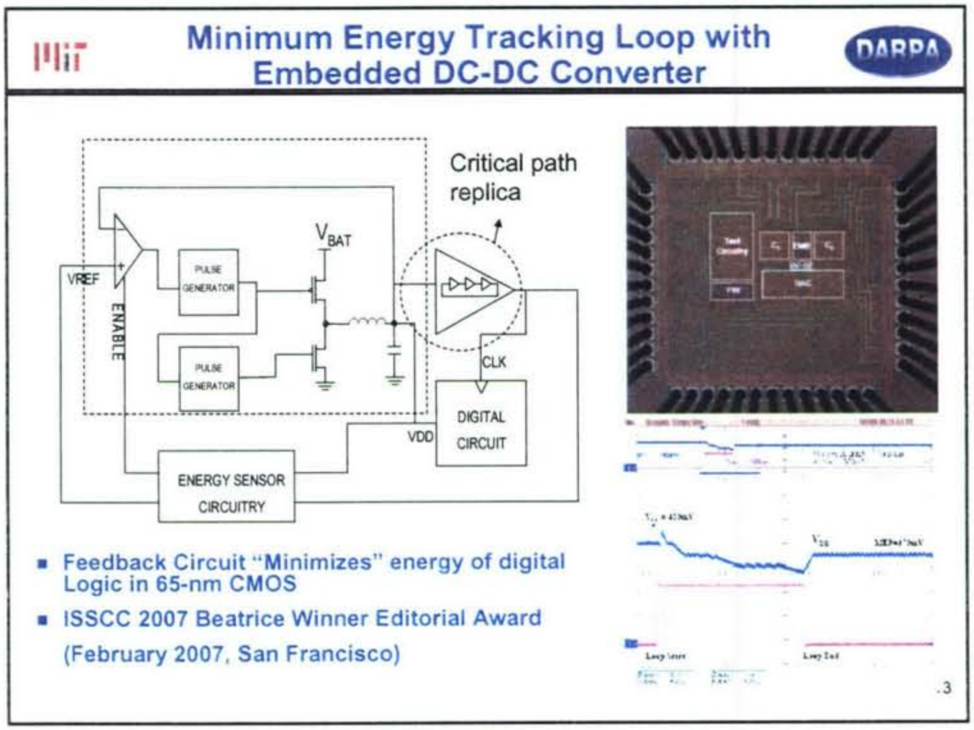
Lowest Operating Voltage SRAM (<350mV) in 65-nm CMOS Demonstrated



Motivation for Energy Minimizing Loop

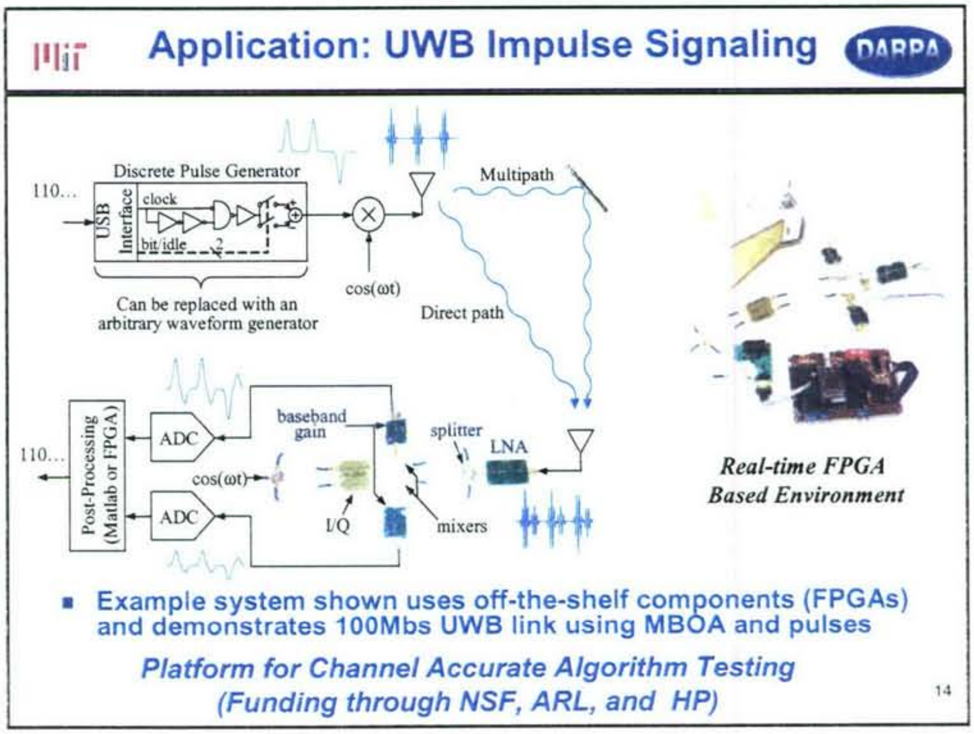
- 7-tap FIR filter
- MEP = 400mV
- 6X savings in energy obtained by operating at the MEP compared to the nominal voltage of 1.2V

- MEP moves with change in workload – no. of taps of the FIR
- A further 2.2X improvement in energy consumed can be obtained by tracking the MEP



Minimum Energy Tracking Loop with Embedded DC-DC Converter

- Feedback Circuit "Minimizes" energy of digital Logic in 65-nm CMOS
- ISSCC 2007 Beatrice Winner Editorial Award (February 2007, San Francisco)

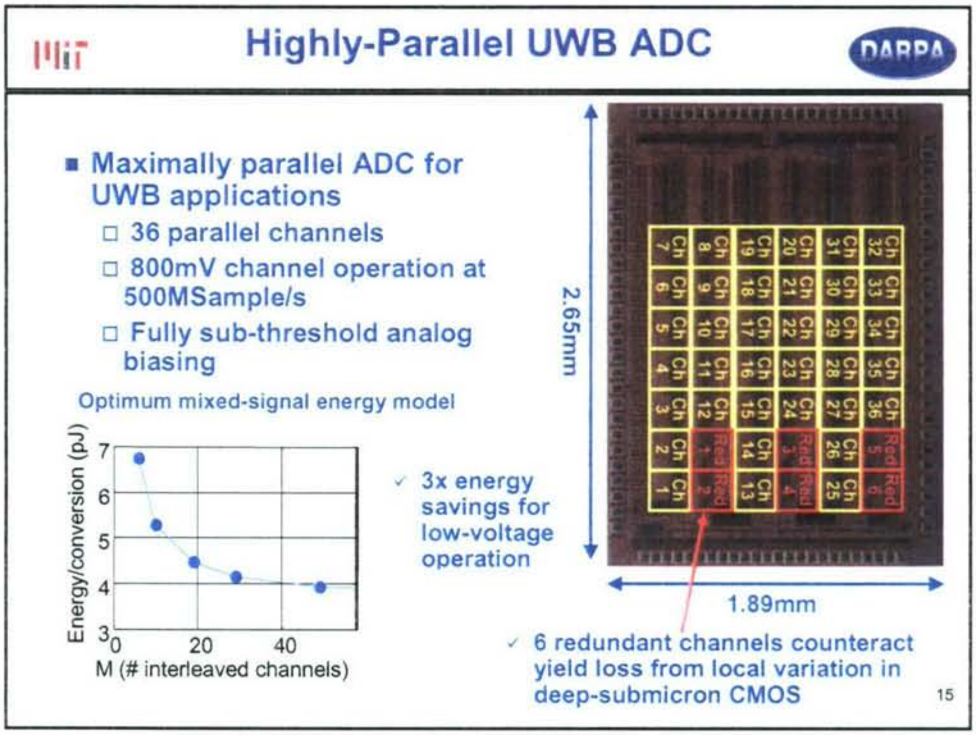


Application: UWB Impulse Signaling

Real-time FPGA Based Environment

- Example system shown uses off-the-shelf components (FPGAs) and demonstrates 100Mbs UWB link using MBOA and pulses

Platform for Channel Accurate Algorithm Testing (Funding through NSF, ARL, and HP)



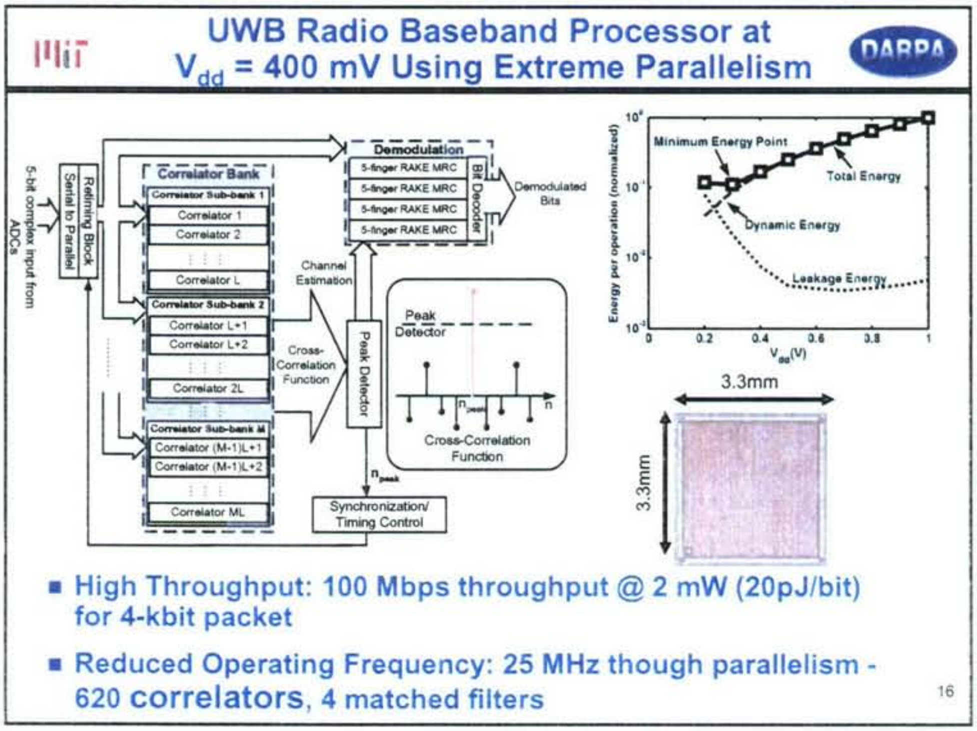
Highly-Parallel UWB ADC

- Maximally parallel ADC for UWB applications
  - □ 36 parallel channels
  - □ 800mV channel operation at 500MSample/s
  - □ Fully sub-threshold analog biasing

Optimum mixed-signal energy model

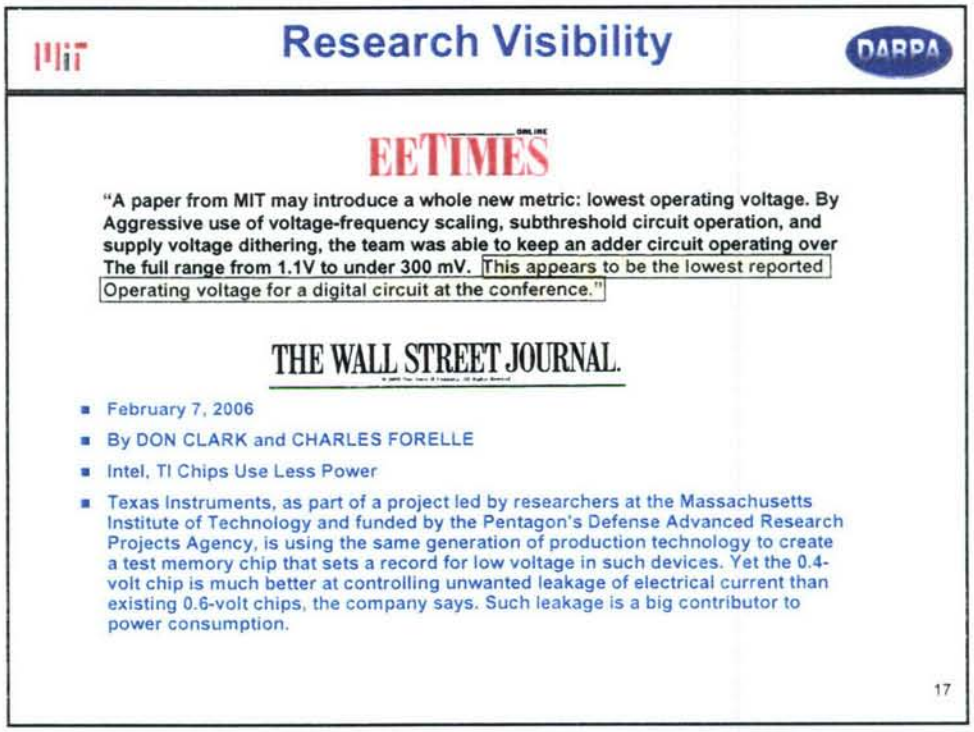
✓ 3x energy savings for low-voltage operation

✓ 6 redundant channels counteract yield loss from local variation in deep-submicron CMOS



UWB Radio Baseband Processor at V_{dd} = 400 mV Using Extreme Parallelism

- High Throughput: 100 Mbps throughput @ 2 mW (20pJ/bit) for 4-kbit packet
- Reduced Operating Frequency: 25 MHz though parallelism - 620 correlators, 4 matched filters



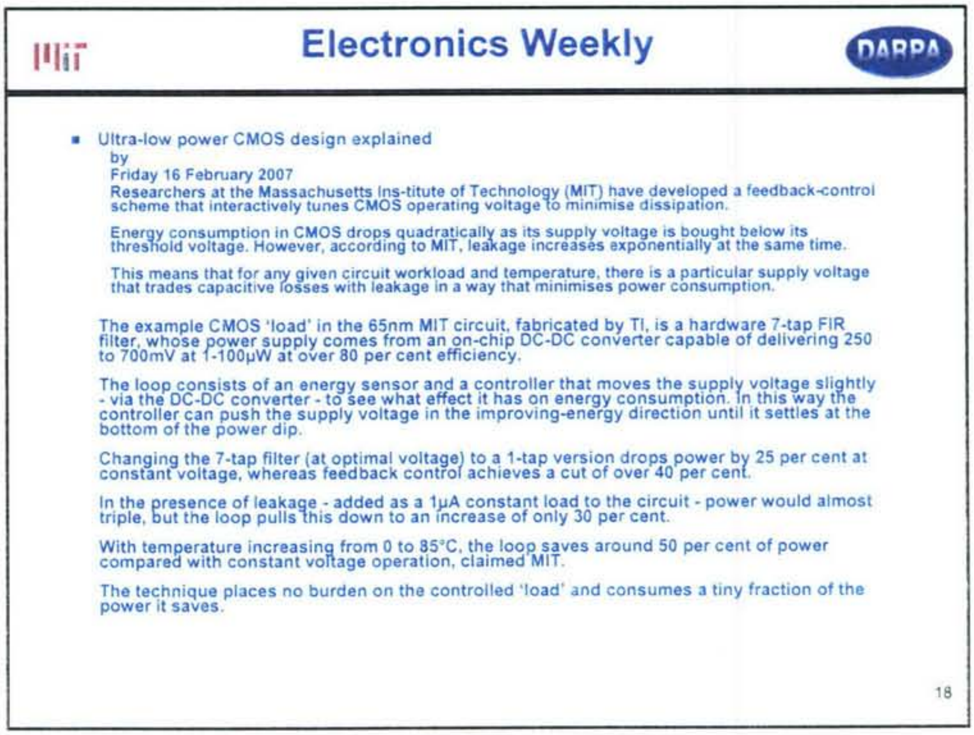
Research Visibility

EETIMES

"A paper from MIT may introduce a whole new metric: lowest operating voltage. By Aggressive use of voltage-frequency scaling, subthreshold circuit operation, and supply voltage dithering, the team was able to keep an adder circuit operating over The full range from 1.1V to under 300 mV. This appears to be the lowest reported Operating voltage for a digital circuit at the conference."

THE WALL STREET JOURNAL.

- February 7, 2006
- By DON CLARK and CHARLES FORELLE
- Intel, TI Chips Use Less Power
- Texas Instruments, as part of a project led by researchers at the Massachusetts Institute of Technology and funded by the Pentagon's Defense Advanced Research Projects Agency, is using the same generation of production technology to create a test memory chip that sets a record for low voltage in such devices. Yet the 0.4-volt chip is much better at controlling unwanted leakage of electrical current than existing 0.6-volt chips, the company says. Such leakage is a big contributor to power consumption.



Electronics Weekly

- Ultra-low power CMOS design explained

  by
  Friday 16 February 2007
  Researchers at the Massachusetts Ins-titute of Technology (MIT) have developed a feedback-control scheme that interactively tunes CMOS operating voltage to minimise dissipation.

  Energy consumption in CMOS drops quadratically as its supply voltage is bought below its threshold voltage. However, according to MIT, leakage increases exponentially at the same time.

  This means that for any given circuit workload and temperature, there is a particular supply voltage that trades capacitive losses with leakage in a way that minimises power consumption.

The example CMOS 'load' in the 65nm MIT circuit, fabricated by TI, is a hardware 7-tap FIR filter, whose power supply comes from an on-chip DC-DC converter capable of delivering 250 to 700mV at 1-100µW at over 80 per cent efficiency.

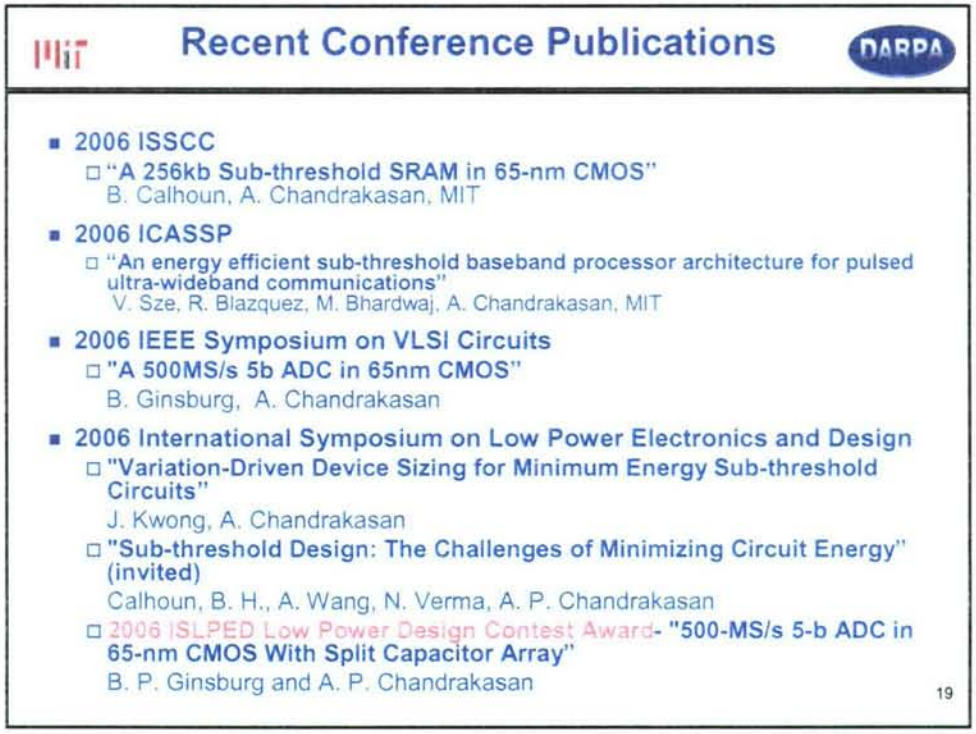
The loop consists of an energy sensor and a controller that moves the supply voltage slightly - via the DC-DC converter - to see what effect it has on energy consumption. In this way the controller can push the supply voltage in the improving-energy direction until it settles at the bottom of the power dip.

Changing the 7-tap filter (at optimal voltage) to a 1-tap version drops power by 25 per cent at constant voltage, whereas feedback control achieves a cut of over 40 per cent.

In the presence of leakage - added as a 1µA constant load to the circuit - power would almost triple, but the loop pulls this down to an increase of only 30 per cent.

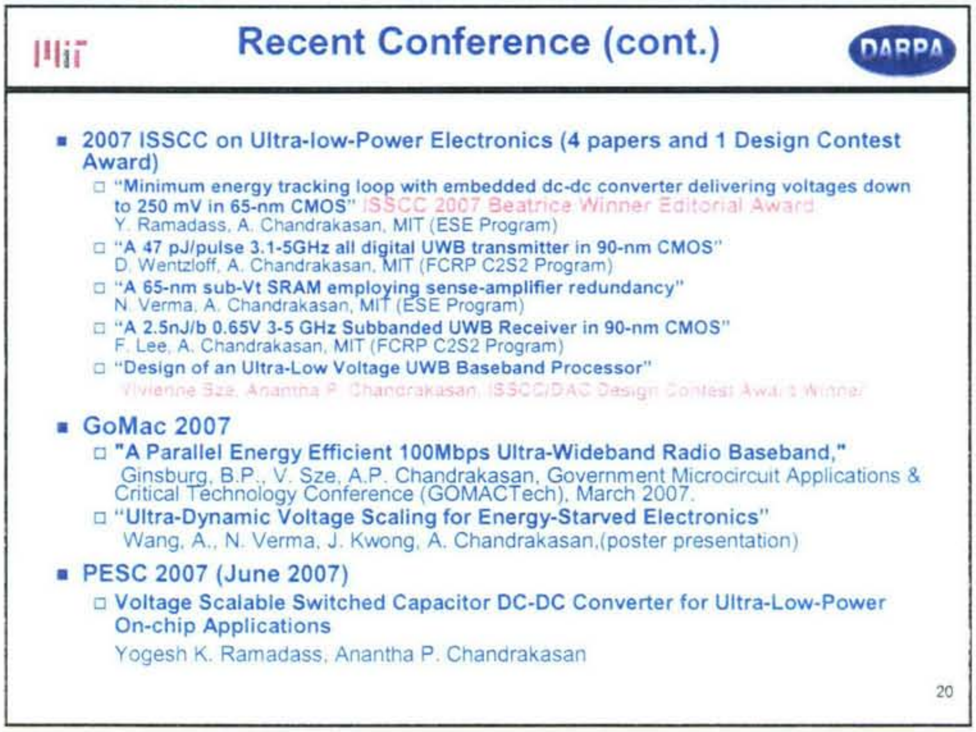
With temperature increasing from 0 to 85°C, the loop saves around 50 per cent of power compared with constant voltage operation, claimed MIT.

The technique places no burden on the controlled 'load' and consumes a tiny fraction of the power it saves.



Recent Conference Publications

- 2006 ISSCC
  - □ "A 256kb Sub-threshold SRAM in 65-nm CMOS"
    B. Calhoun, A. Chandrakasan, MIT
- 2006 ICASSP
  - □ "An energy efficient sub-threshold baseband processor architecture for pulsed ultra-wideband communications"
    V. Sze, R. Blazquez, M. Bhardwaj, A. Chandrakasan, MIT
- 2006 IEEE Symposium on VLSI Circuits
  - □ "A 500MS/s 5b ADC in 65nm CMOS"
    B. Ginsburg, A. Chandrakasan
- 2006 International Symposium on Low Power Electronics and Design
  - □ "Variation-Driven Device Sizing for Minimum Energy Sub-threshold Circuits"
    J. Kwong, A. Chandrakasan
  - □ "Sub-threshold Design: The Challenges of Minimizing Circuit Energy" (invited)
    Calhoun, B. H., A. Wang, N. Verma, A. P. Chandrakasan
  - □ 2006 ISLPED Low Power Design Contest Award- "500-MS/s 5-b ADC in 65-nm CMOS With Split Capacitor Array"
    B. P. Ginsburg and A. P. Chandrakasan



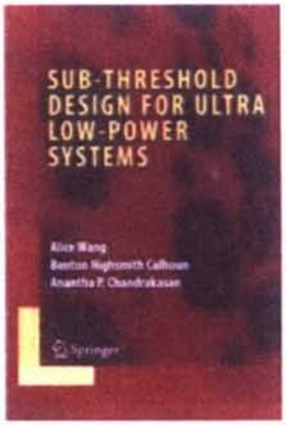
Recent Conference (cont.)

- 2007 ISSCC on Ultra-low-Power Electronics (4 papers and 1 Design Contest Award)
  - □ "Minimum energy tracking loop with embedded dc-dc converter delivering voltages down to 250 mV in 65-nm CMOS" ISSCC 2007 Beatrice Winner Editorial Award
    Y. Ramadass, A. Chandrakasan, MIT (ESE Program)
  - □ "A 47 pJ/pulse 3.1-5GHz all digital UWB transmitter in 90-nm CMOS"
    D. Wentzloff, A. Chandrakasan, MIT (FCRP C2S2 Program)
  - □ "A 65-nm sub-Vt SRAM employing sense-amplifier redundancy"
    N. Verma, A. Chandrakasan, MIT (ESE Program)
  - □ "A 2.5nJ/b 0.65V 3-5 GHz Subbanded UWB Receiver in 90-nm CMOS"
    F. Lee, A. Chandrakasan, MIT (FCRP C2S2 Program)
  - □ "Design of an Ultra-Low Voltage UWB Baseband Processor"
    Vivienne Sze, Anantha P. Chandrakasan, ISSCC/DAC Design Contest Award Winner
- GoMac 2007
  - □ "A Parallel Energy Efficient 100Mbps Ultra-Wideband Radio Baseband,"
    Ginsburg, B.P., V. Sze, A.P. Chandrakasan, Government Microcircuit Applications & Critical Technology Conference (GOMACTech), March 2007.
  - □ "Ultra-Dynamic Voltage Scaling for Energy-Starved Electronics"
    Wang, A., N. Verma, J. Kwong, A. Chandrakasan,(poster presentation)
- PESC 2007 (June 2007)
  - □ Voltage Scalable Switched Capacitor DC-DC Converter for Ultra-Low-Power On-chip Applications
    Yogesh K. Ramadass, Anantha P. Chandrakasan

Journal Papers

- Benton H. Calhoun, Anantha P. Chandrakasan, "Static Noise Margin Variation for Sub-threshold SRAM in 65nm CMOS," IEEE Journal of Solid-State Circuits, vol. 41, no. 7, pp. 1673-1679, July 2006.
- Benton H. Calhoun, Anantha P. Chandrakasan, "A 256kb 65nm Sub-threshold SRAM Design for Ultra-Low Voltage Operation", IEEE Journal of Solid-State Circuits, pp. 680-688, March 2007.
- B. P. Ginsburg and A. P. Chandrakasan, "500-MS/s 5-bit ADC in 65-nm CMOS With Split Capacitor Array DAC," IEEE J. Solid-State Circuits, vol. 42, no. 4, pp. 739-747, Apr. 2007.
- Invited (to the special issue of the JSSC):
  - □ N. Verma, A. P. Chandrakasan, "A 256kb 65nm 8T Sub-Threshold SRAM employing Sense-Amplifier Redundancy"
  - □ R. Yogesh, A. P. Chandrakasan, "A minimum energy tracking loop with embedded DC-DC Converter enabling ultra-low-voltage operation in 65nm CMOS"

Book on Sub-threshold Circuits

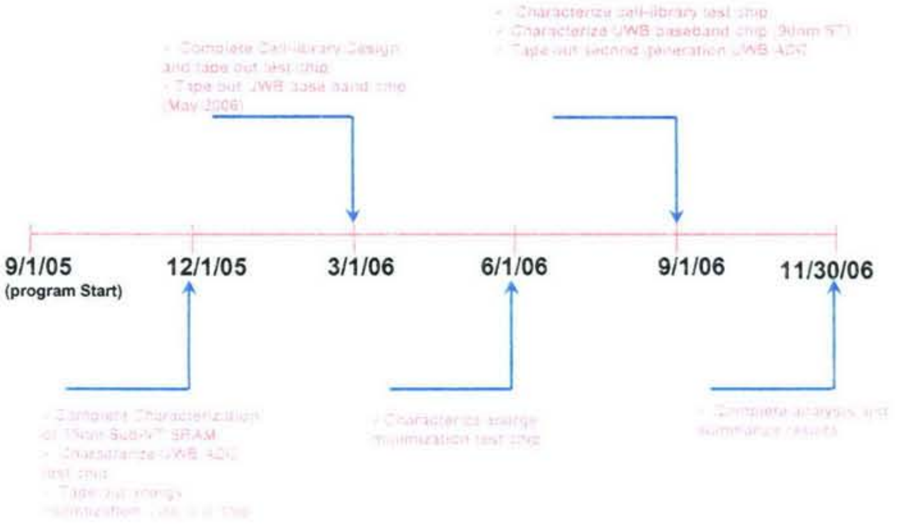


- A Direct output of research from the DARPA ESE and Seedling efforts
- Also includes invited chapters from Eric Vittoz (pioneer of sub-threshold Analog circuits) and Christian Enz (EKV model)

Impact and Technology Transitions

- Boeing exploring asynchronous sub-threshold logic
  - □ Transferred our Verilog design and sizing methodology (Joyce Kwong and Vivienne Sze)
- ISSCC and ISLPED has many submissions on sub-threshold logic design
- Implementation of sub-threshold MSP-430 (in collaboration with TI) on-going with standard cell library, sub-VT SRAM and DC-DC converters

Schedule and Milestones



- Complete Cell-library Design and tape out test chip
- Tape out UWB base band chip (May 2006)

- Characterize cell-library test chip
- Characterize UWB baseband chip (90nm ST)
- Tape out second generation UWB ADC

- Complete Characterization of 65nm Sub-VT SRAM
- Characterize UWB ADC test chip
- Tape out energy minimization [illegible] test chip

- Characterize energy minimization test chip

- Complete analysis and summarize results

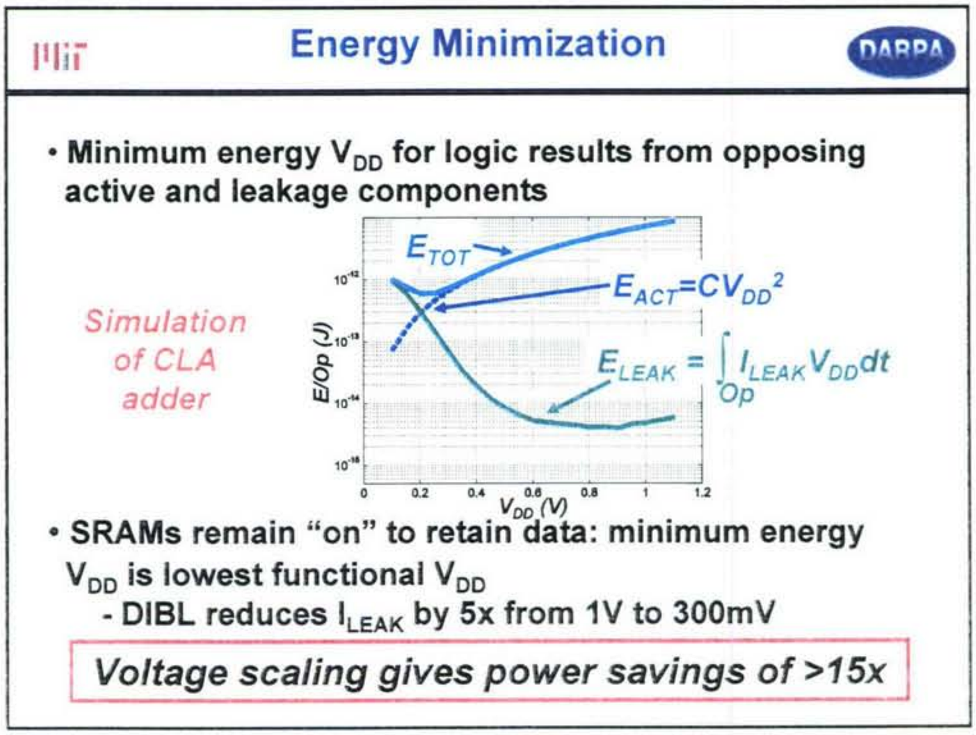
- New result: preliminary design of fully on-chip DC-DC converter

Dense Sub-V_t SRAM For Ultra-Low Leakage Power and Access Energy

PI: Anantha Chandrakasan
(anantha@mtl.mit.edu)

Students: *Naveen Verma, Benton Highsmith Calhoun*

TI Collaborators: Dennis Buss, Terence Breedijk, Uming Ko, David Scott, Dr. Alice Wang



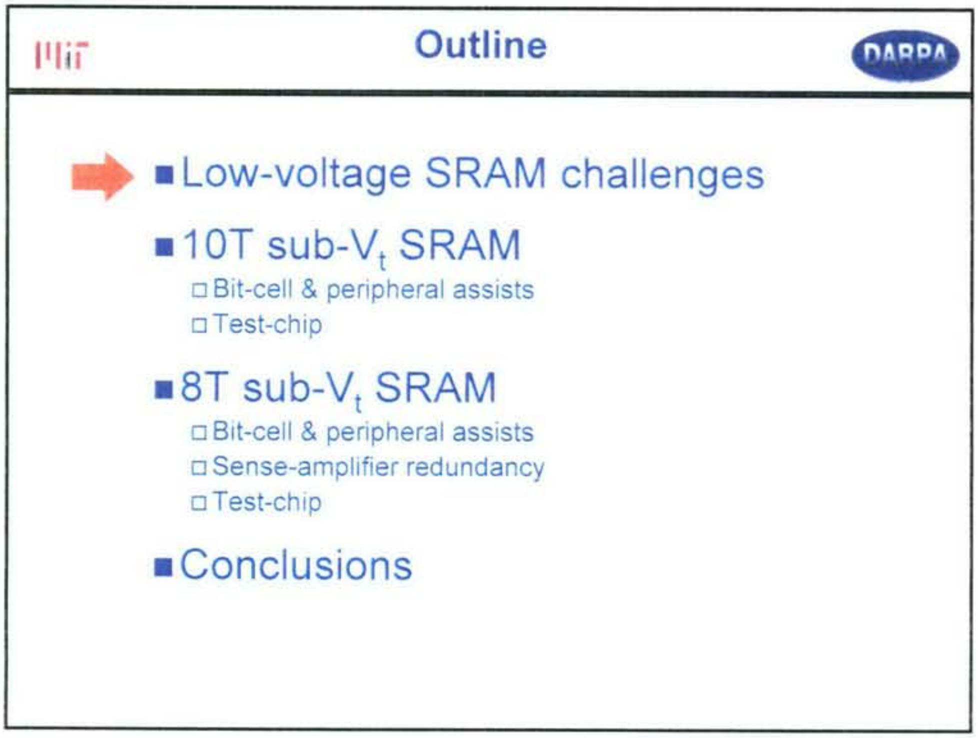
Energy Minimization

- Minimum energy V_{DD} for logic results from opposing active and leakage components

Simulation of CLA adder

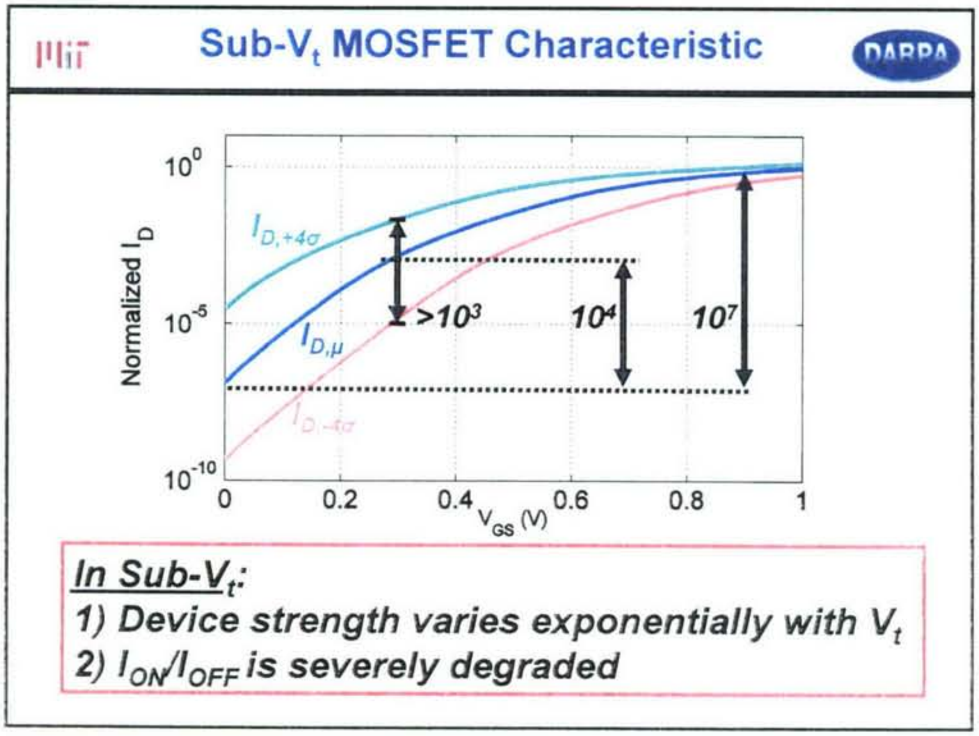
- SRAMs remain "on" to retain data: minimum energy V_{DD} is lowest functional V_{DD}
  - DIBL reduces I_{LEAK} by 5x from 1V to 300mV

Voltage scaling gives power savings of >15x



Outline

- Low-voltage SRAM challenges
- 10T sub-V_t SRAM
  - □ Bit-cell & peripheral assists
  - □ Test-chip
- 8T sub-V_t SRAM
  - □ Bit-cell & peripheral assists
  - □ Sense-amplifier redundancy
  - □ Test-chip
- Conclusions

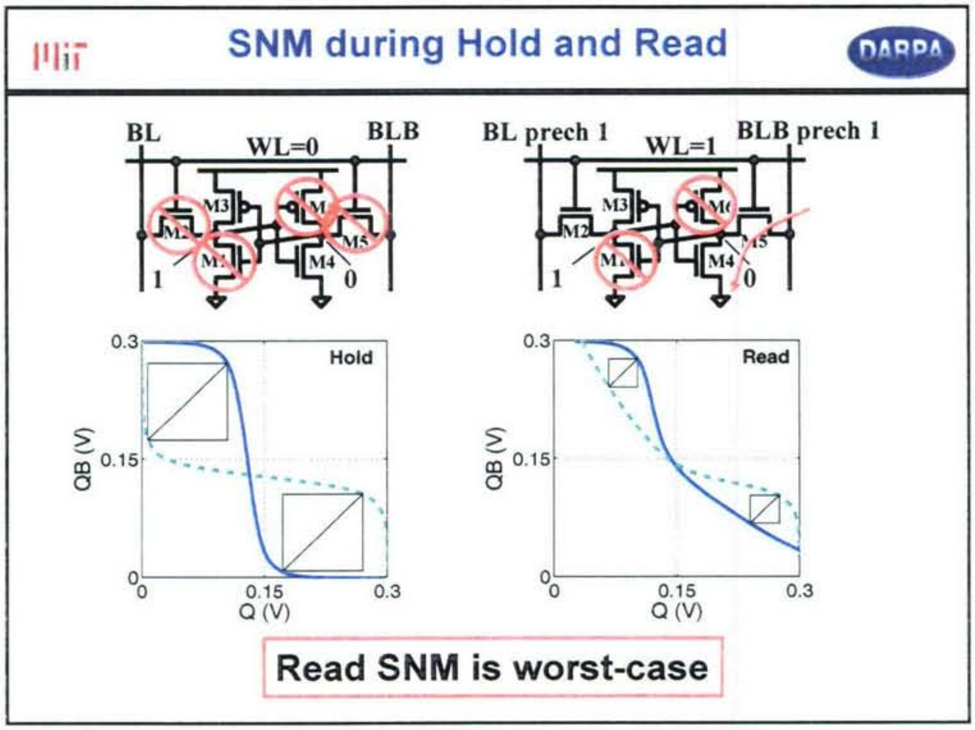


Sub-V_t MOSFET Characteristic

In Sub-V_t:

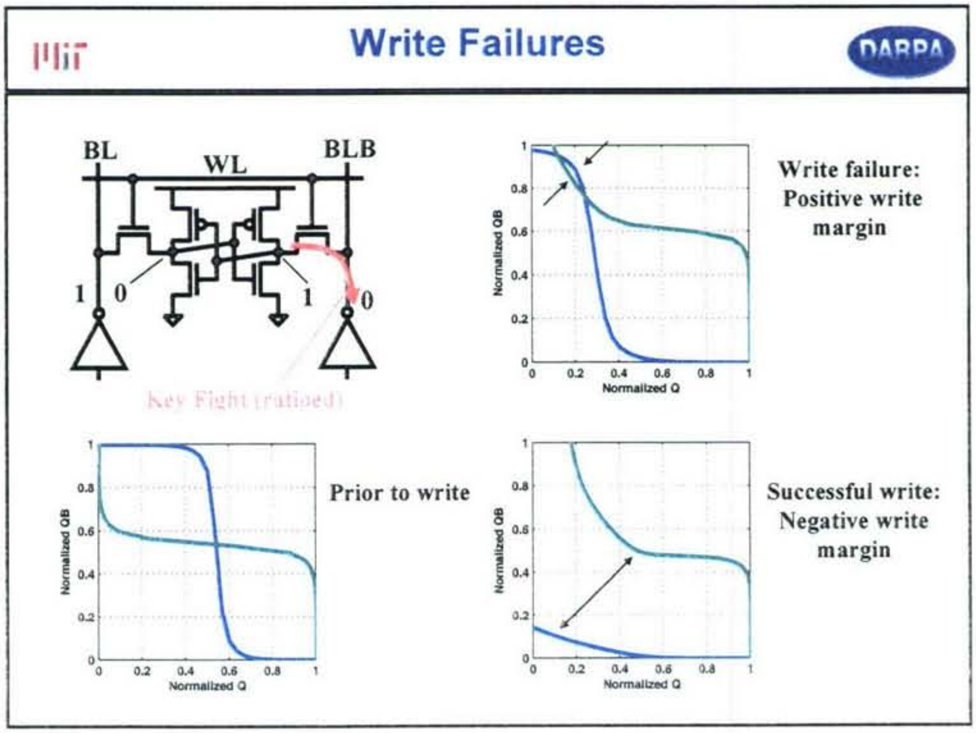
1) Device strength varies exponentially with V_t

2) I_{ON}/I_{OFF} is severely degraded



SNM during Hold and Read

Read SNM is worst-case

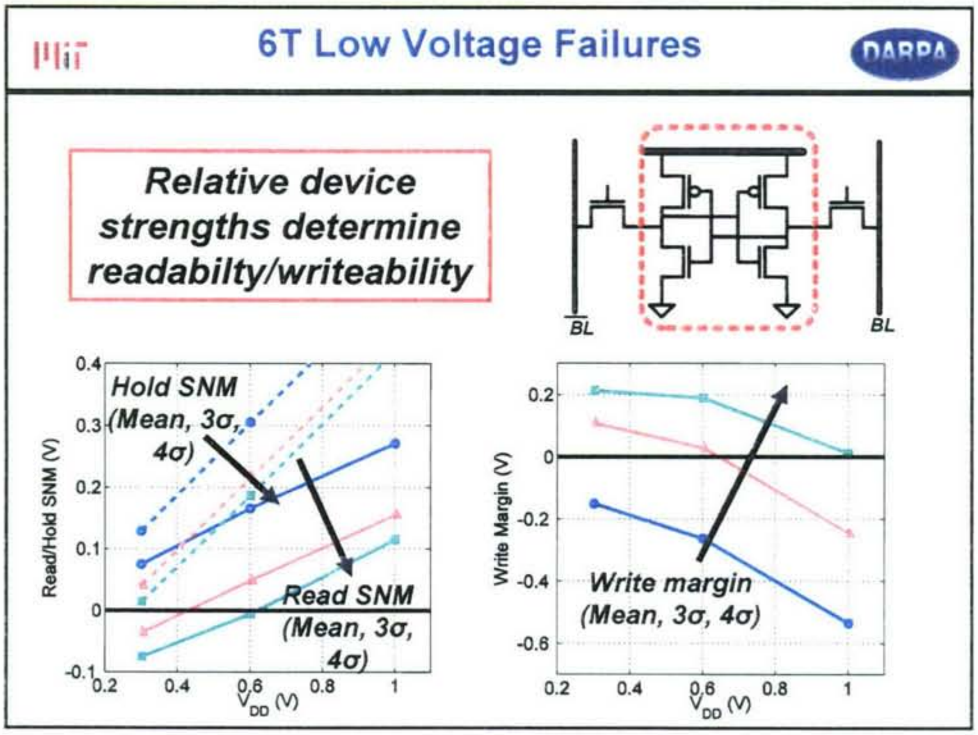


Write Failures

Write failure: Positive write margin

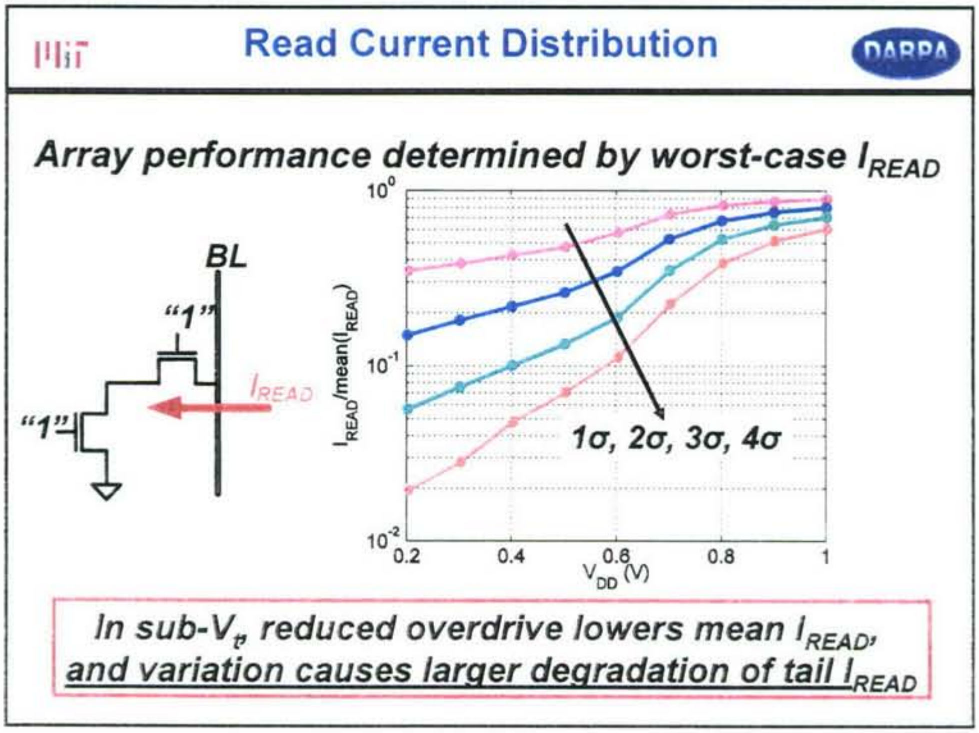
Prior to write

Successful write: Negative write margin



6T Low Voltage Failures

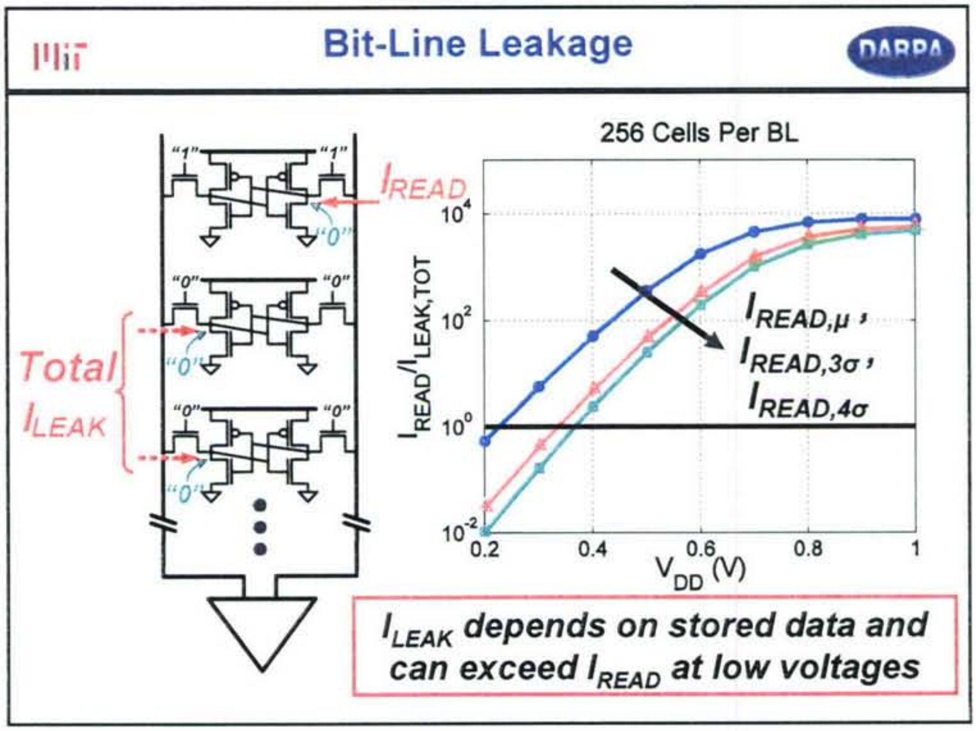
Relative device strengths determine readabilty/writeability



Read Current Distribution

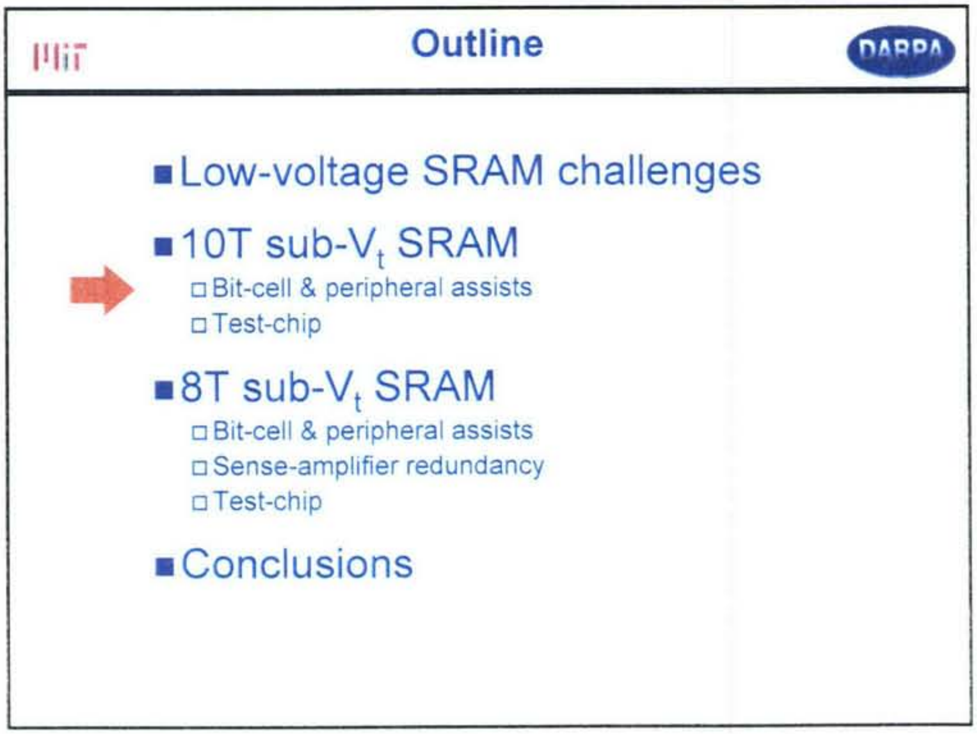
Array performance determined by worst-case I_{READ}

In sub-V_t, reduced overdrive lowers mean I_{READ}, and variation causes larger degradation of tail I_{READ}



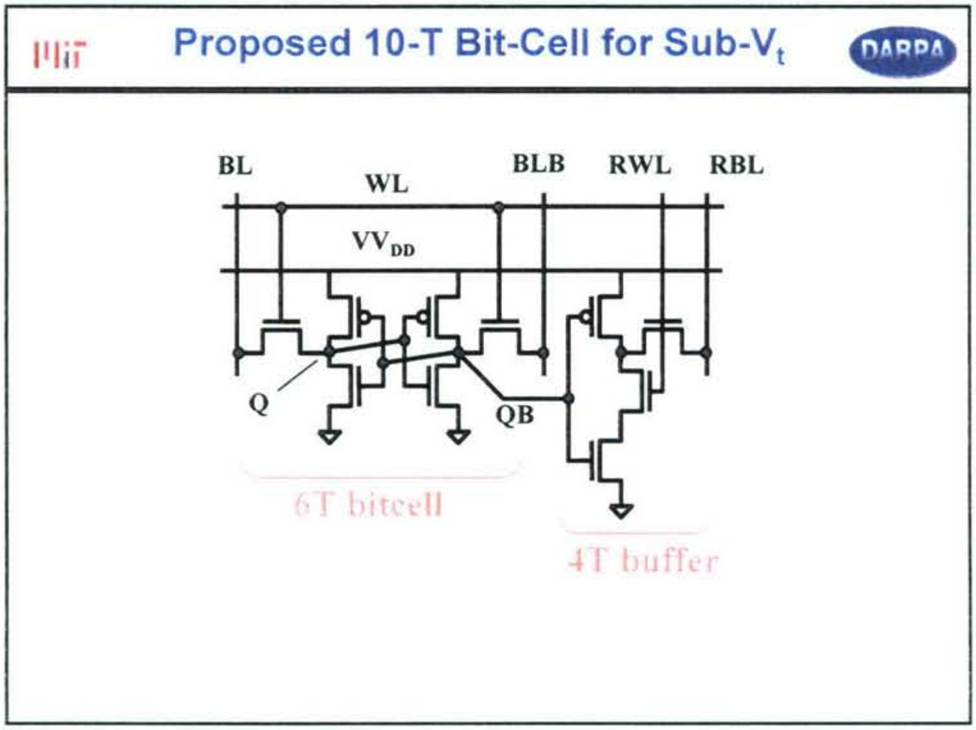
Bit-Line Leakage

I_{LEAK} depends on stored data and can exceed I_{READ} at low voltages

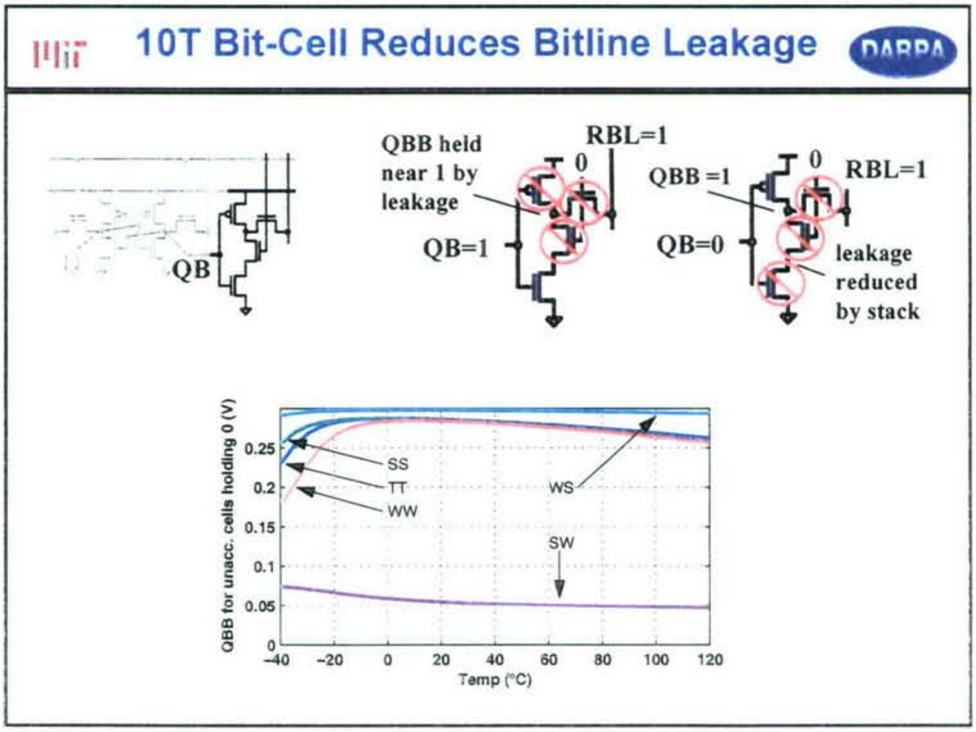


Outline

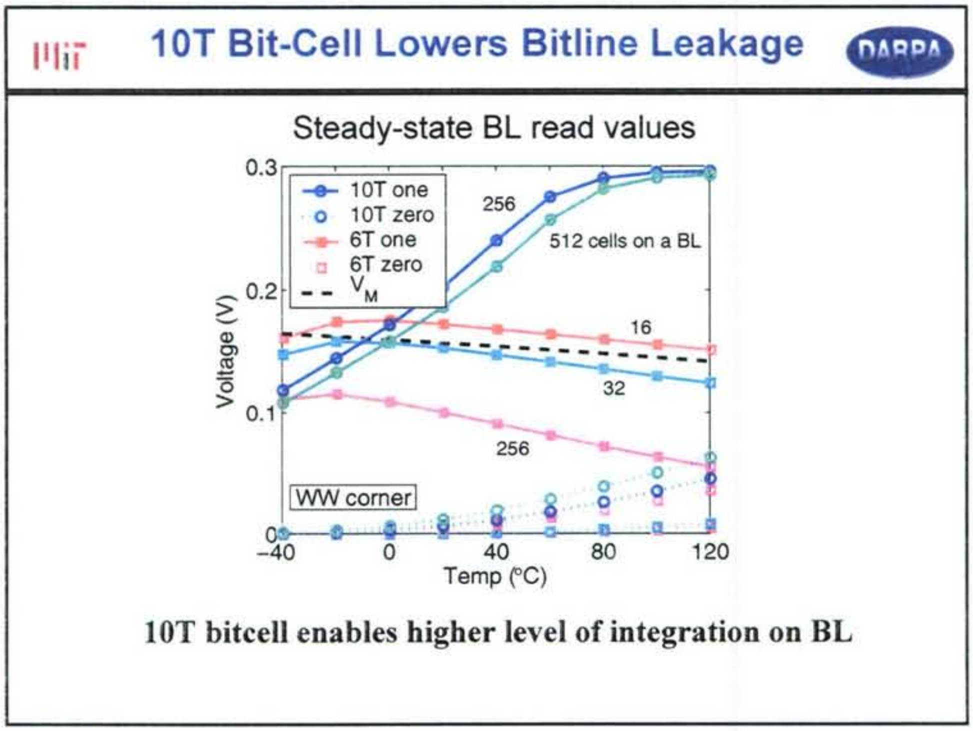
- Low-voltage SRAM challenges
- 10T sub-V_t SRAM
  - □ Bit-cell & peripheral assists
  - □ Test-chip
- 8T sub-V_t SRAM
  - □ Bit-cell & peripheral assists
  - □ Sense-amplifier redundancy
  - □ Test-chip
- Conclusions



Proposed 10-T Bit-Cell for Sub-V_t



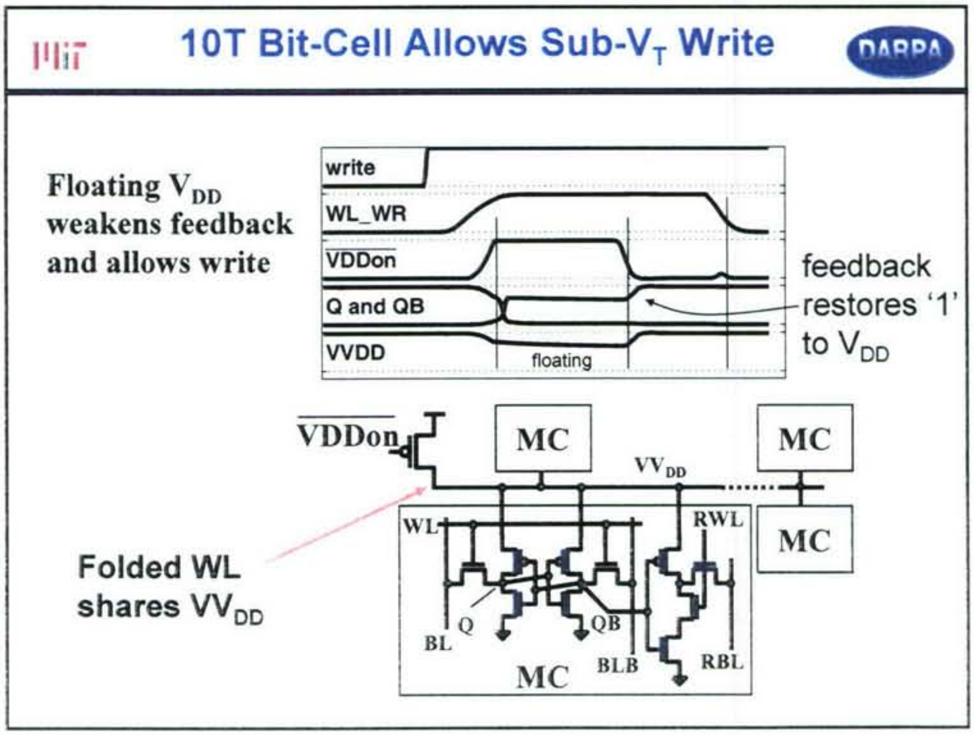
10T Bit-Cell Reduces Bitline Leakage



10T Bit-Cell Lowers Bitline Leakage

Steady-state BL read values

10T bitcell enables higher level of integration on BL



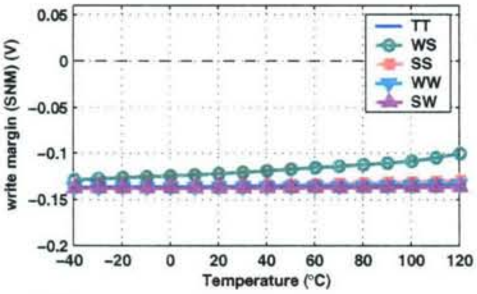
10T Bit-Cell Allows Sub-V_T Write

Floating V_{DD} weakens feedback and allows write

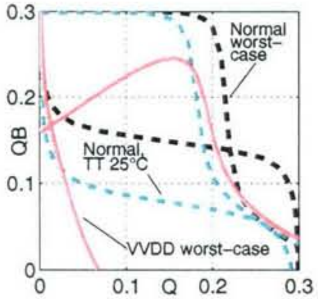
feedback restores '1' to V_{DD}

Folded WL shares VV_{DD}

10T Bit-Cell Allows Sub-V_T Write



Write successful across process corner and temperature

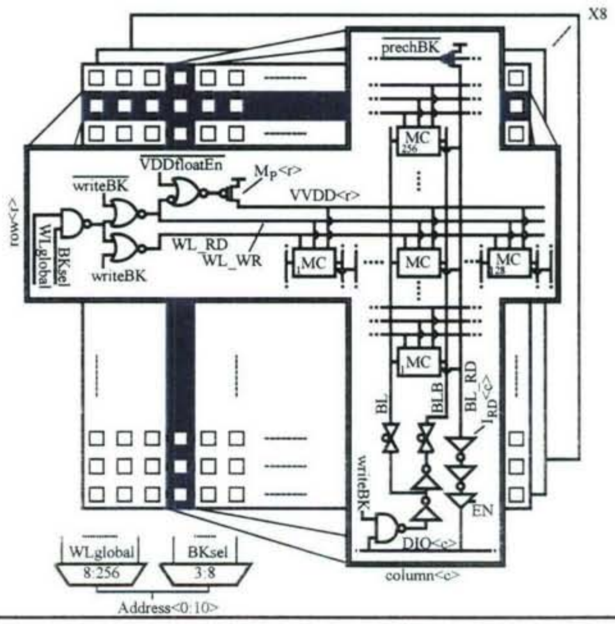


Write margin correctly negative at worst-case corner

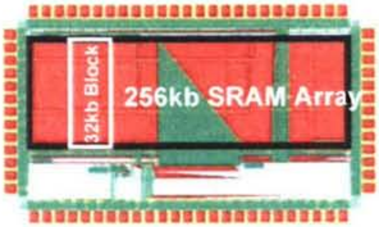
Test Chip Architecture

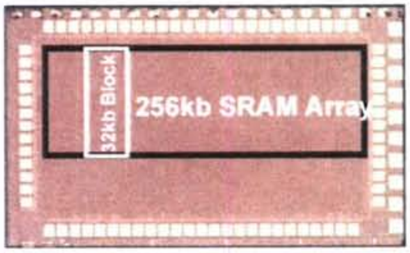
- 256 rows and 128 columns per block
- Static CMOS peripherals
- Separate WL V_{DD} for boosting
- Assumed 1x1 redundancy
- Simulation:

Operates at 300mV across all process corners from 0 to 100°C



256Kb 65nm Sub-V_T memory



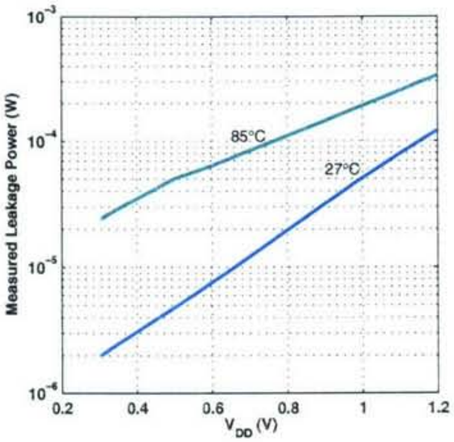


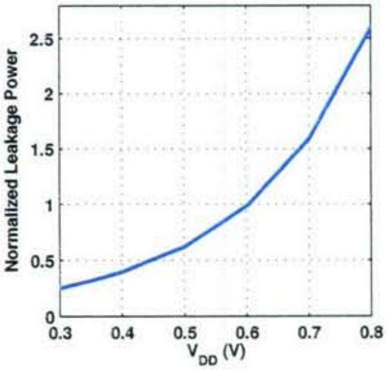
Test chip addressing the sub-V_T problems using 10T bitcell: 1.89mm by 1.12mm

Chip functions to below 400mV, holds without error to <250mV:
At 400mV, 3.28μW and 475kHz at 27°C

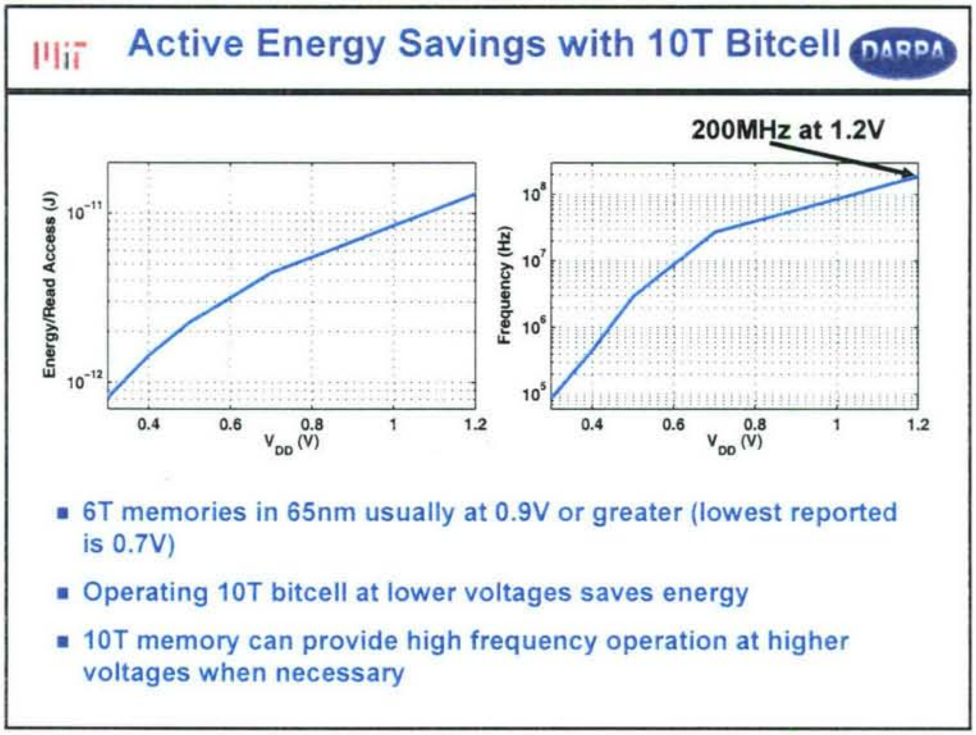
Reads without error to <u>320mV</u> (27°C) and <u>360mV</u> (85°C)
Write without error to <u>380mV</u> (27°C) and <u>350mV</u> (85°C)

Power Measurements



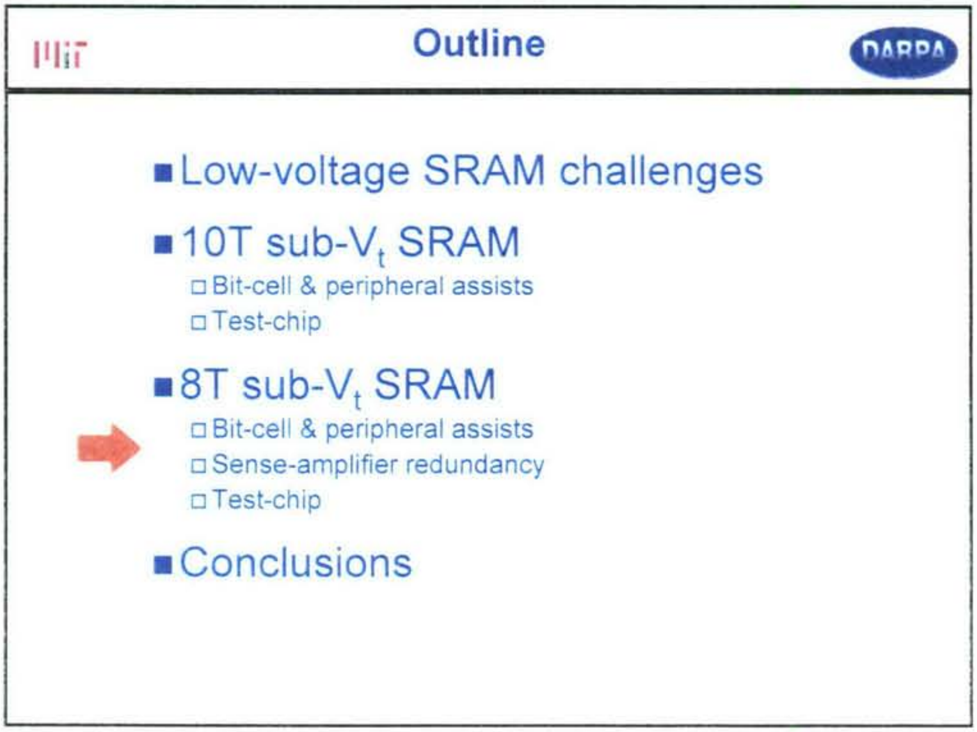


Relative to 0.6V 6T SRAM, 2.2X less leakage power at 0.4V and
3.3X less leakage power at 0.3V
>60X less leakage power than 1.2V



Active Energy Savings with 10T Bitcell

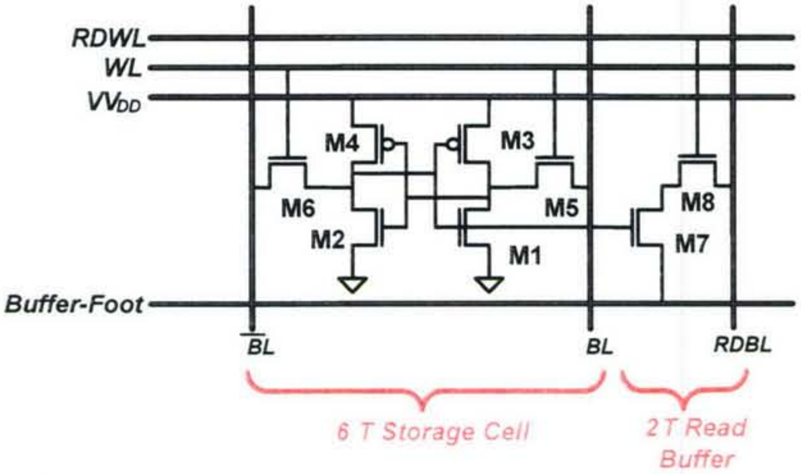
- 6T memories in 65nm usually at 0.9V or greater (lowest reported is 0.7V)
- Operating 10T bitcell at lower voltages saves energy
- 10T memory can provide high frequency operation at higher voltages when necessary



Outline

- Low-voltage SRAM challenges
- 10T sub-V_t SRAM
  - □ Bit-cell & peripheral assists
  - □ Test-chip
- 8T sub-V_t SRAM
  - □ Bit-cell & peripheral assists
  - □ Sense-amplifier redundancy
  - □ Test-chip
- Conclusions

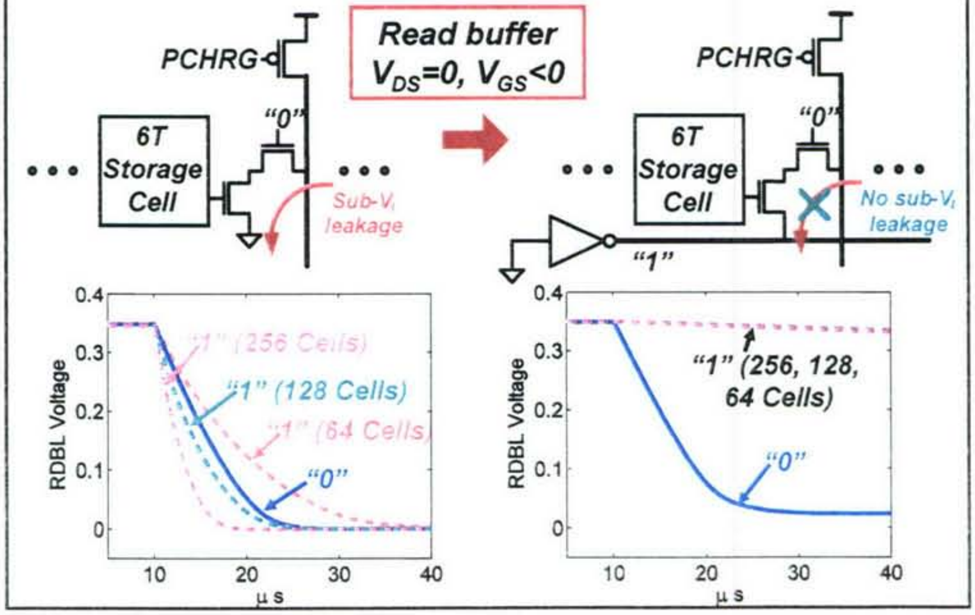
8T Bit-Cell For Sub-V_t SRAM

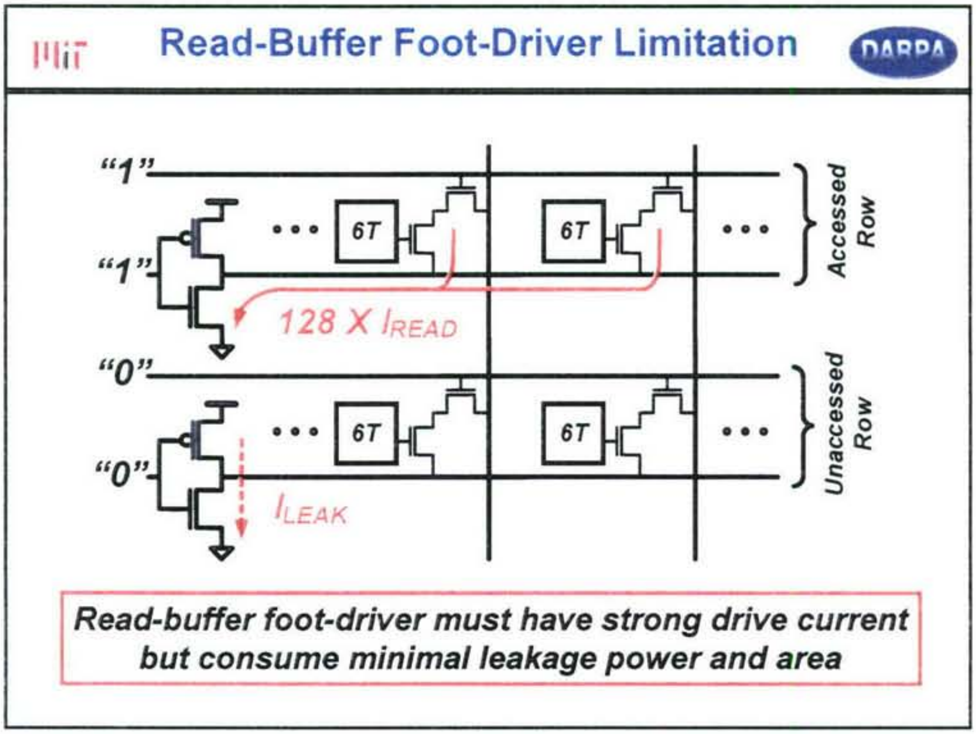


Buffer eliminates read SNM limitation, peripheral assists allow sub-V_t write and sensing

"Zero" Sub-V_t Leakage Read-Buffer

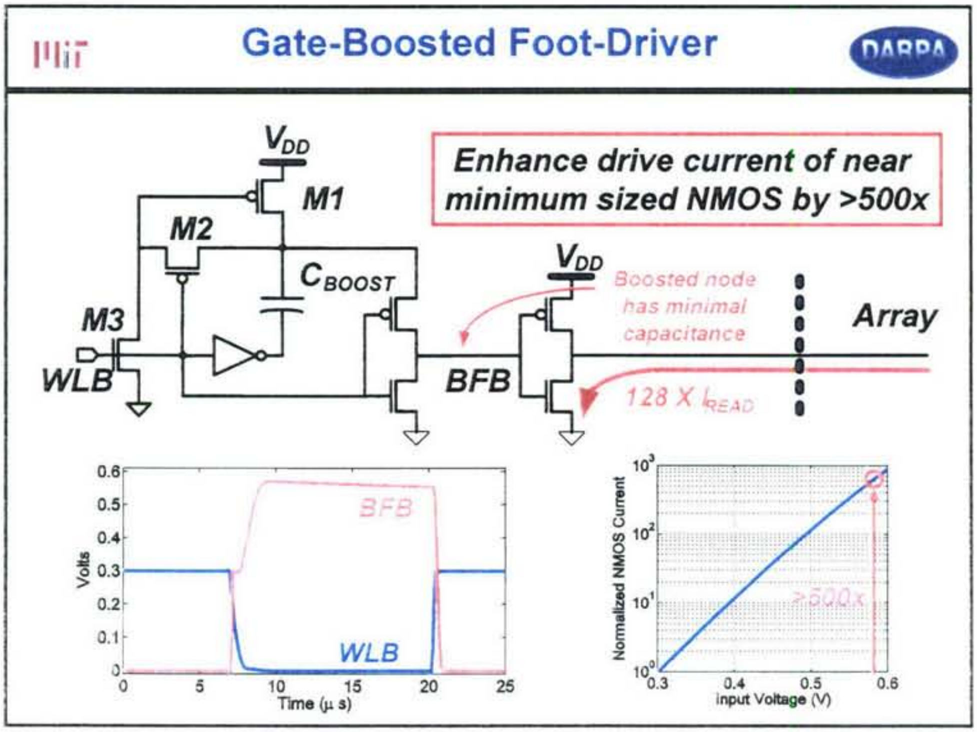
Read buffer
V_{DS}=0, V_{GS}<0





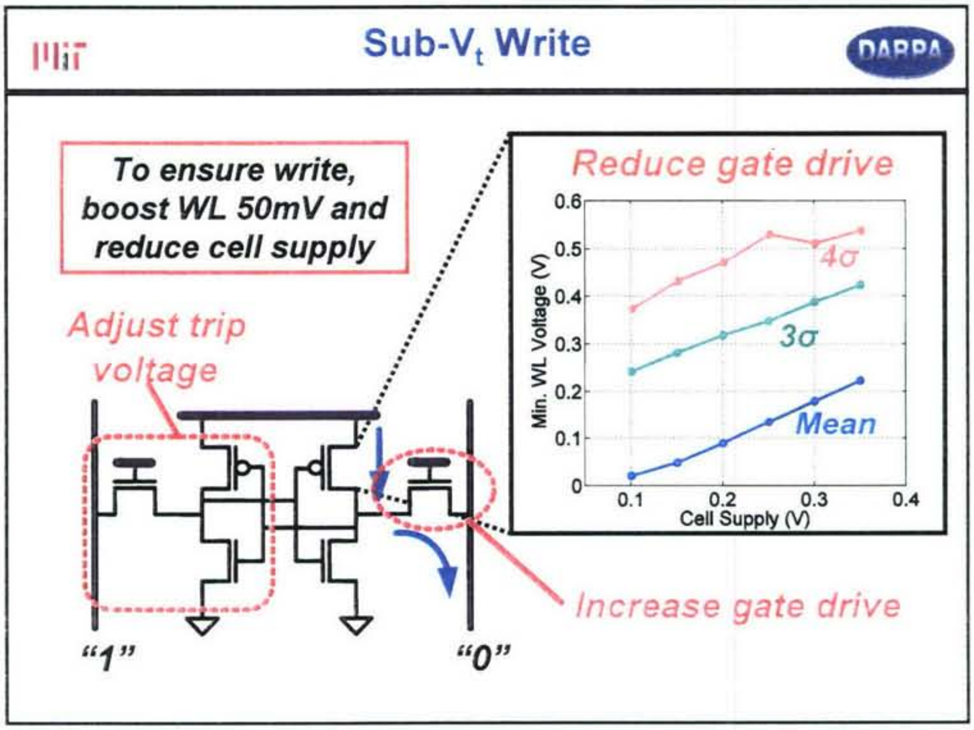
Read-Buffer Foot-Driver Limitation

Read-buffer foot-driver must have strong drive current but consume minimal leakage power and area



Gate-Boosted Foot-Driver

Enhance drive current of near minimum sized NMOS by >500x



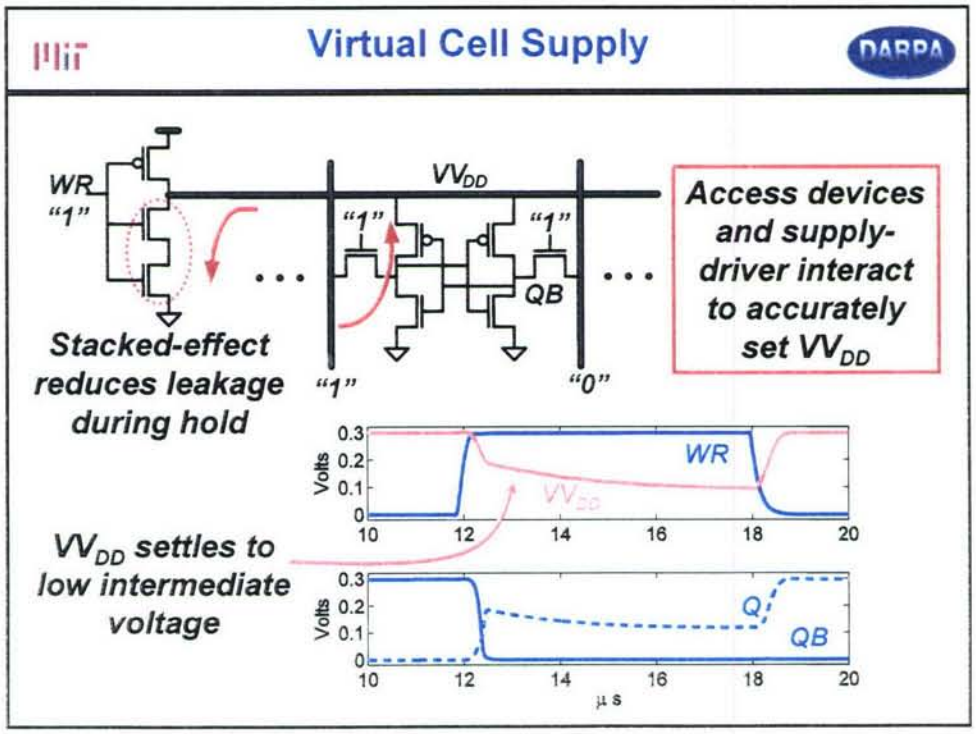
Sub-V_t Write

To ensure write, boost WL 50mV and reduce cell supply

Adjust trip voltage

Reduce gate drive

Increase gate drive



Virtual Cell Supply

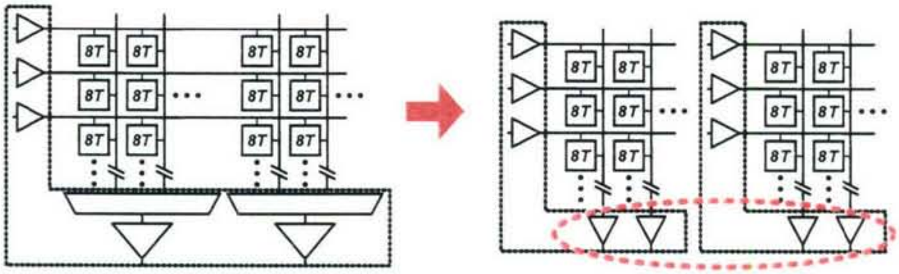
Access devices and supply-driver interact to accurately set VV_{DD}

Stacked-effect reduces leakage during hold

VV_{DD} settles to low intermediate voltage

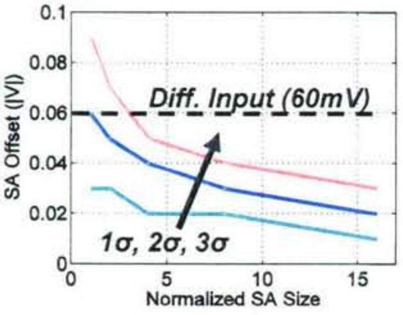
Sense-Amp Area Limitation

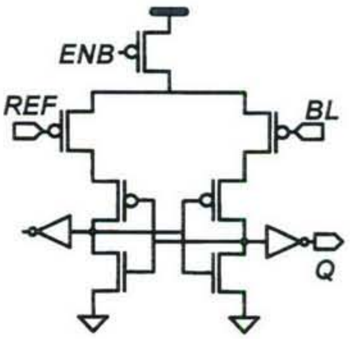
Separate VV_{DD} in non-interleaved layout for minimum voltage and reduced WL load



Sense-amplifier limits yield due to increased number and reduced area

Sense-Amplifier Offset



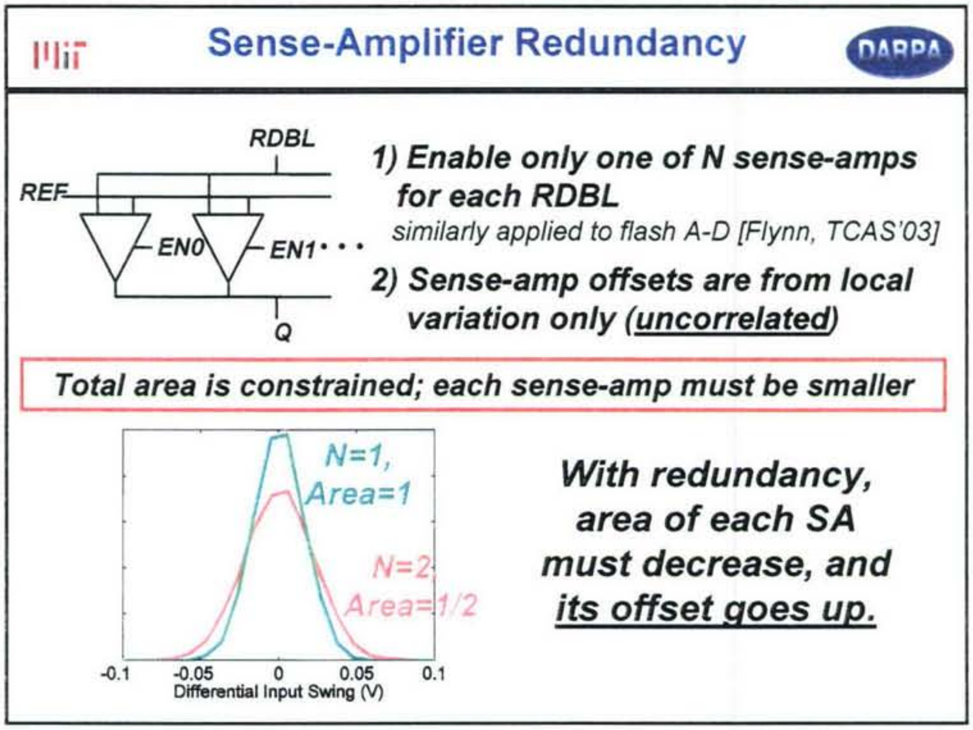


1) **Global variation degrades sense-amp accuracy for single-ended read**

   - Pseudo-differential structure eliminates offset

2) **Local variation results in uncorrelated error distribution of sense-amps**

   - Only device up-sizing can reduce offset deviation



Sense-Amplifier Redundancy

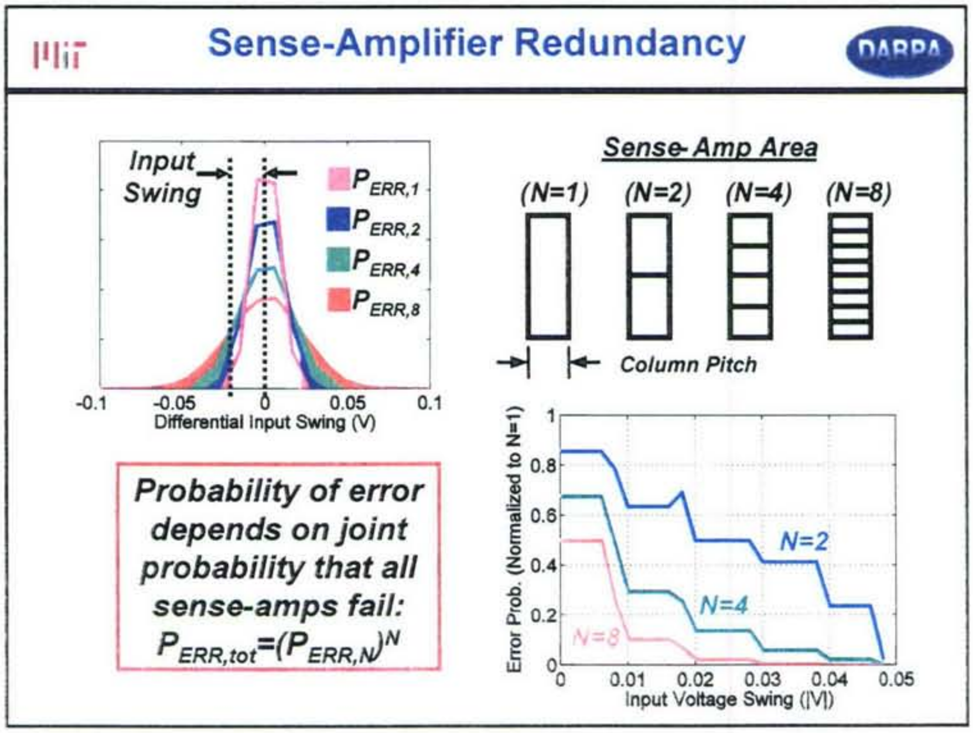
1) Enable only one of N sense-amps for each RDBL

similarly applied to flash A-D [Flynn, TCAS'03]

2) Sense-amp offsets are from local variation only (uncorrelated)

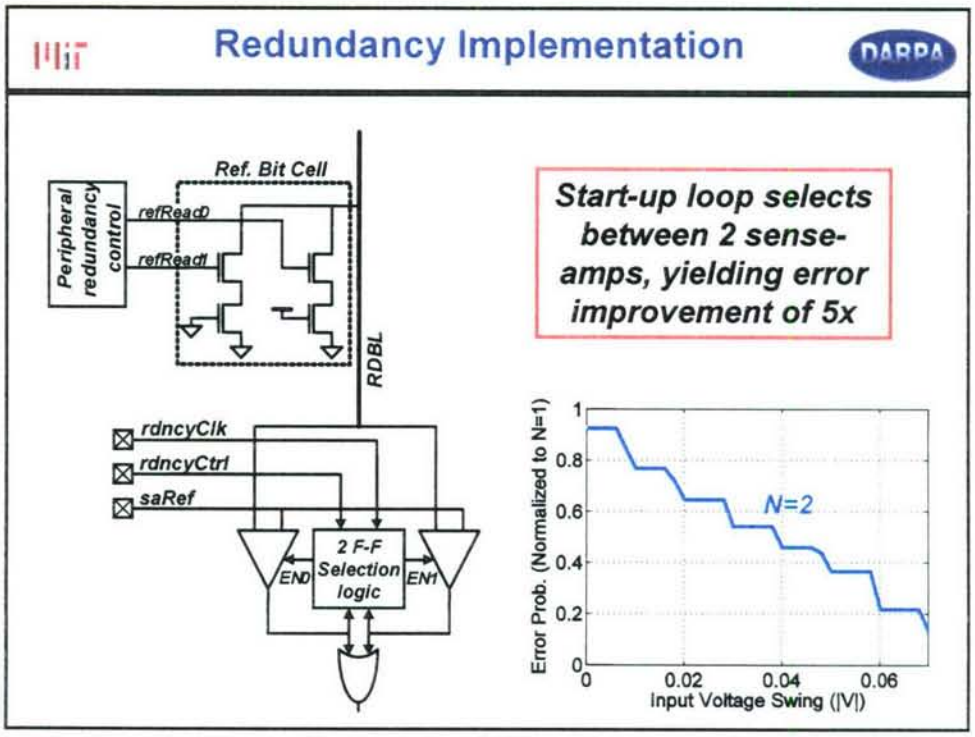
Total area is constrained; each sense-amp must be smaller

With redundancy, area of each SA must decrease, and its offset goes up.



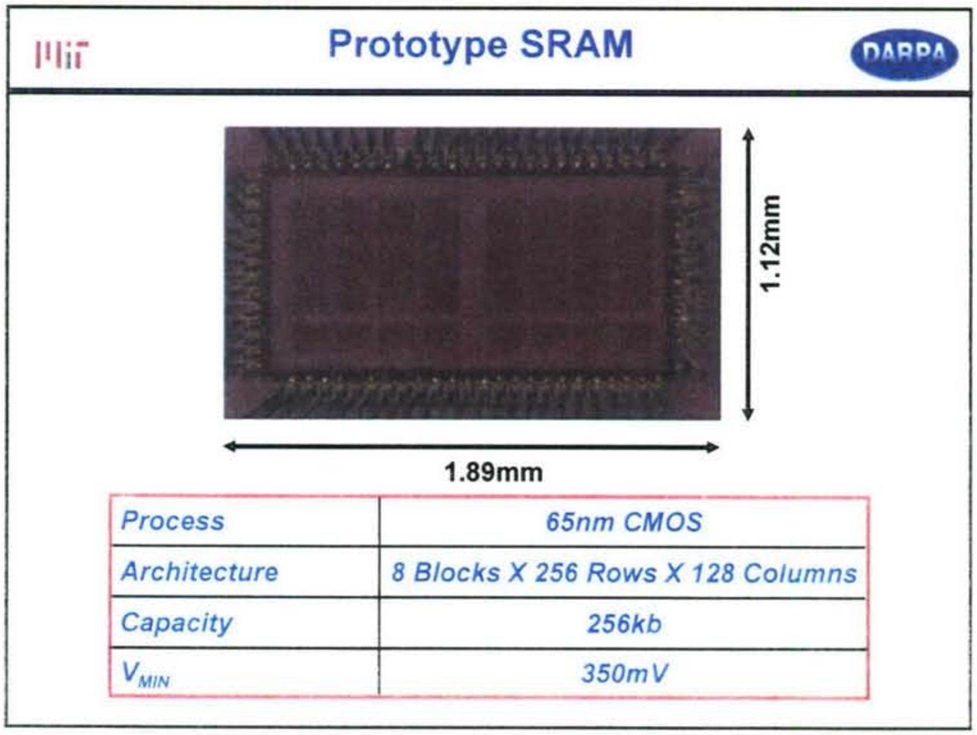
Sense-Amplifier Redundancy

Probability of error depends on joint probability that all sense-amps fail: $P_{ERR,tot}=(P_{ERR,N})^N$



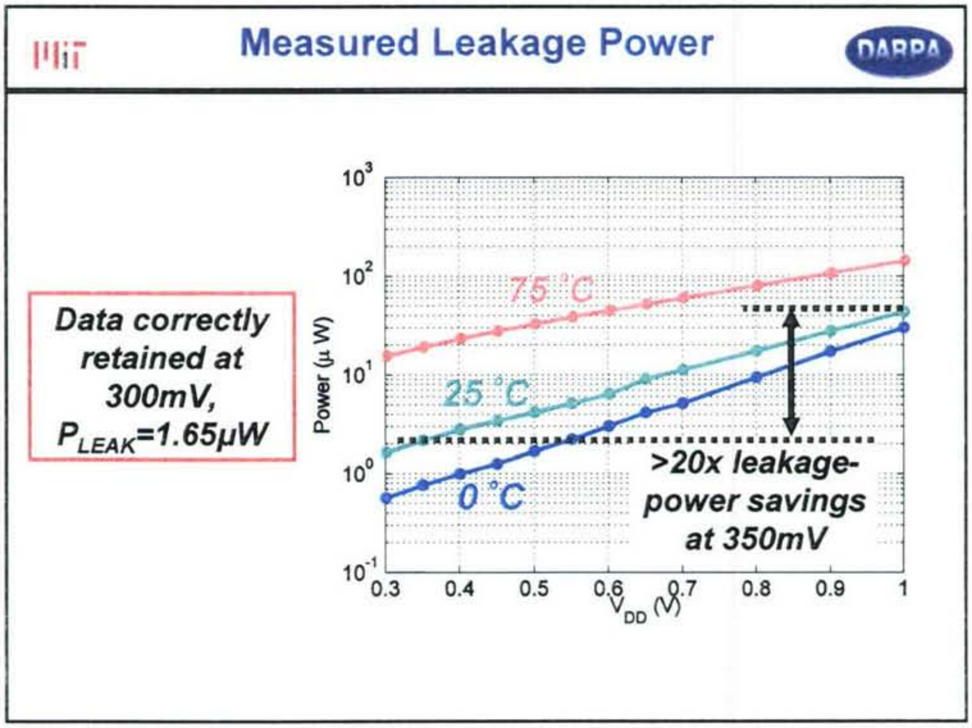
Redundancy Implementation

Start-up loop selects between 2 sense-amps, yielding error improvement of 5x



Prototype SRAM

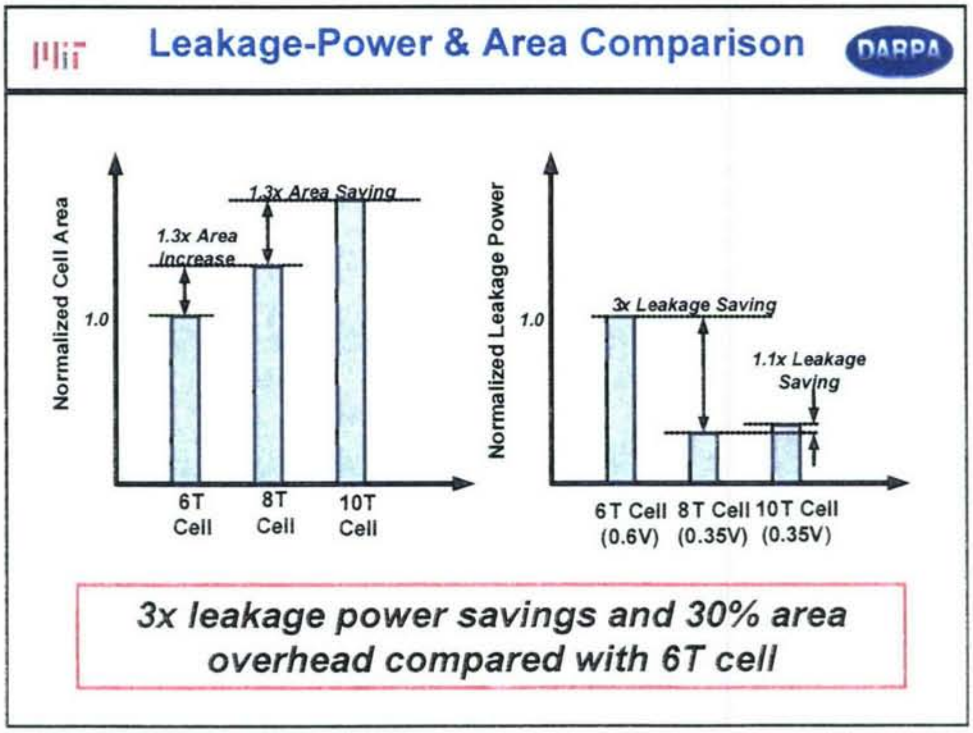
Process	*65nm CMOS*
Architecture	*8 Blocks X 256 Rows X 128 Columns*
Capacity	*256kb*
V_{MIN}	*350mV*



Measured Leakage Power

Data correctly retained at 300mV, P_{LEAK}=1.65μW

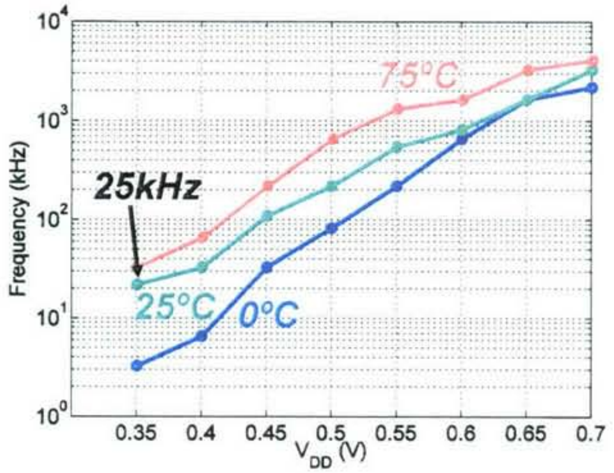
>20x leakage-power savings at 350mV



Leakage-Power & Area Comparison

3x leakage power savings and 30% area overhead compared with 6T cell

Active Performance



Conclusions

- Standard 6T cell is ratioed and limited to ~0.6-0.7V, ~16 cells per bitline
- 10T cell expands read/write margins and manages bit-line leakage in read-buffer increasing robustness to sub-V_t variation
- 8T cell uses peripheral assists to eliminate sub-V_t bit-line leakage and weaken local cell feedback during write
- Sense-amplifier redundancy improves area-offset trade-off by factor of 5
- Sub-V_t bit-cells and peripheral assists allow V_{MIN} to 350mV improving active energy and leakage power by up to factor of 20 compared to conventional 6T limit

Sub-threshold Library Design

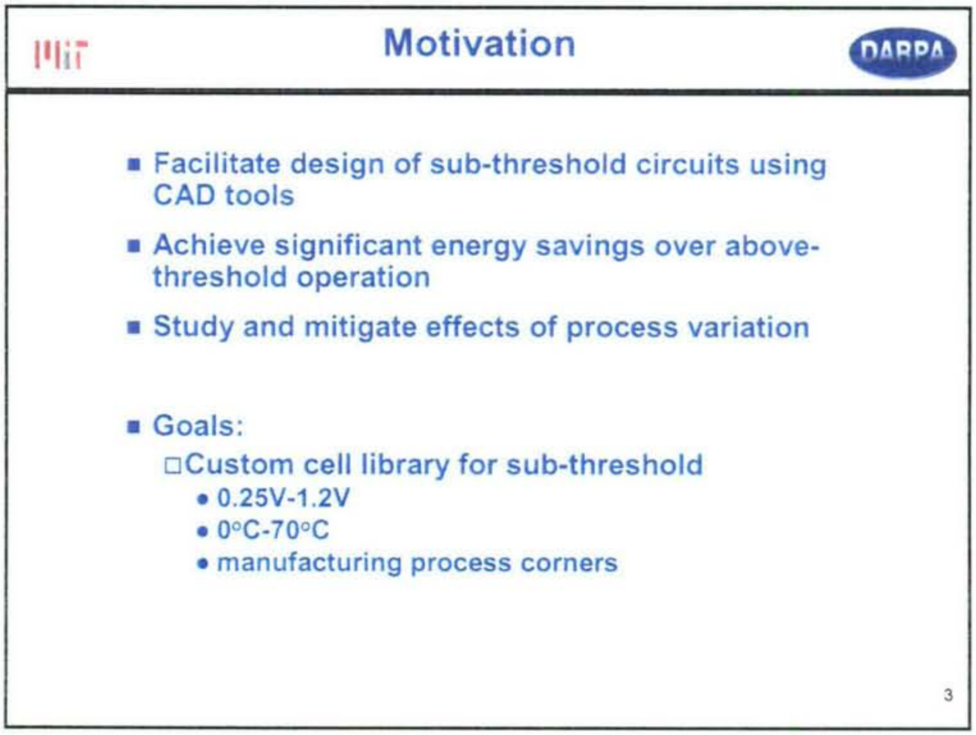
ESE DARPA Program Final Review

April 10, 2007

Joyce Kwong

Outline

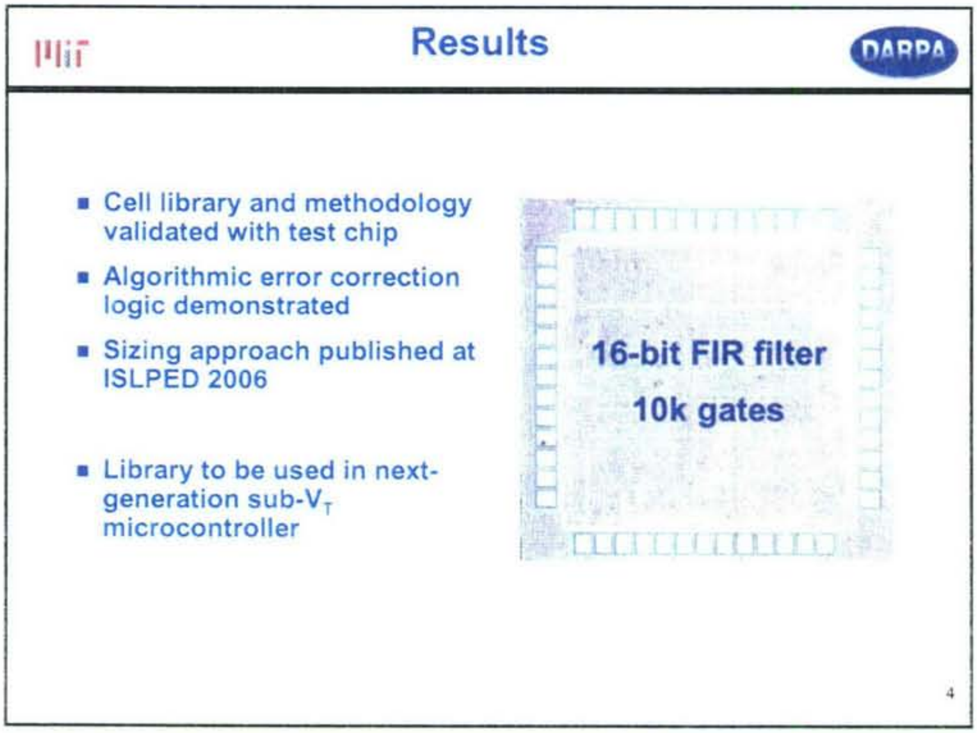
- Motivation
- Results
- Logic design issues in sub-V_T
- Minimum energy operation given yield constraint
- Sub-V_T library test chip
- Algorithmic error correction
- Next phase
  - □ Timing analysis
  - □ Integrated sub-V_T system



Motivation

- Facilitate design of sub-threshold circuits using CAD tools
- Achieve significant energy savings over above-threshold operation
- Study and mitigate effects of process variation

- Goals:
  - □Custom cell library for sub-threshold
    - 0.25V-1.2V
    - 0°C-70°C
    - manufacturing process corners



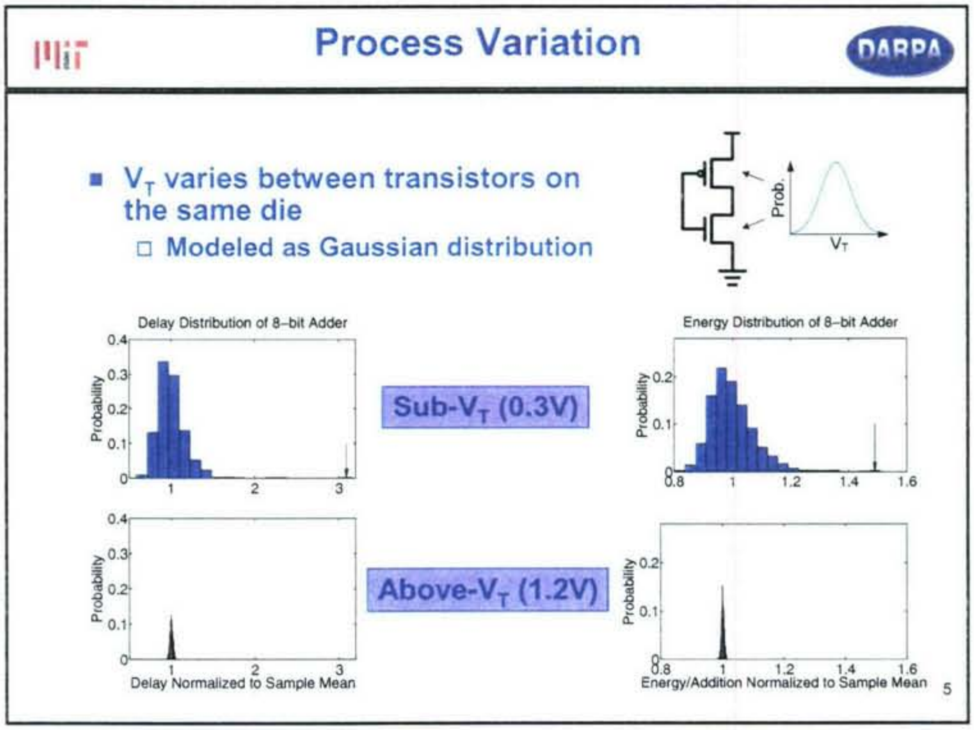
Results

- Cell library and methodology validated with test chip
- Algorithmic error correction logic demonstrated
- Sizing approach published at ISLPED 2006

- Library to be used in next-generation sub-V_T microcontroller

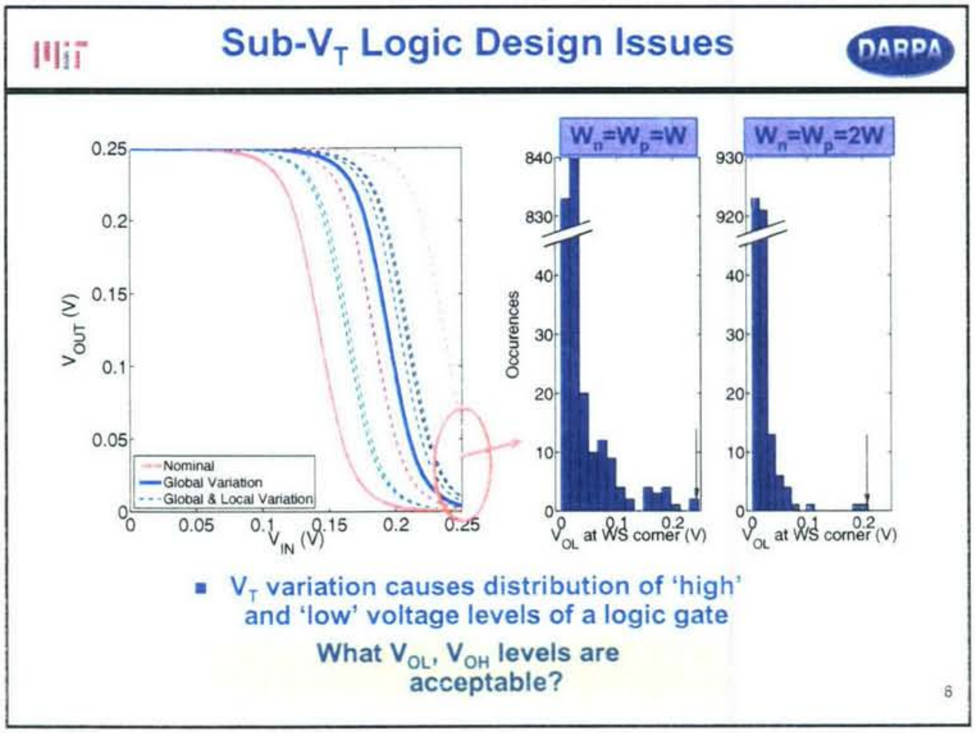
16-bit FIR filter

10k gates



Process Variation

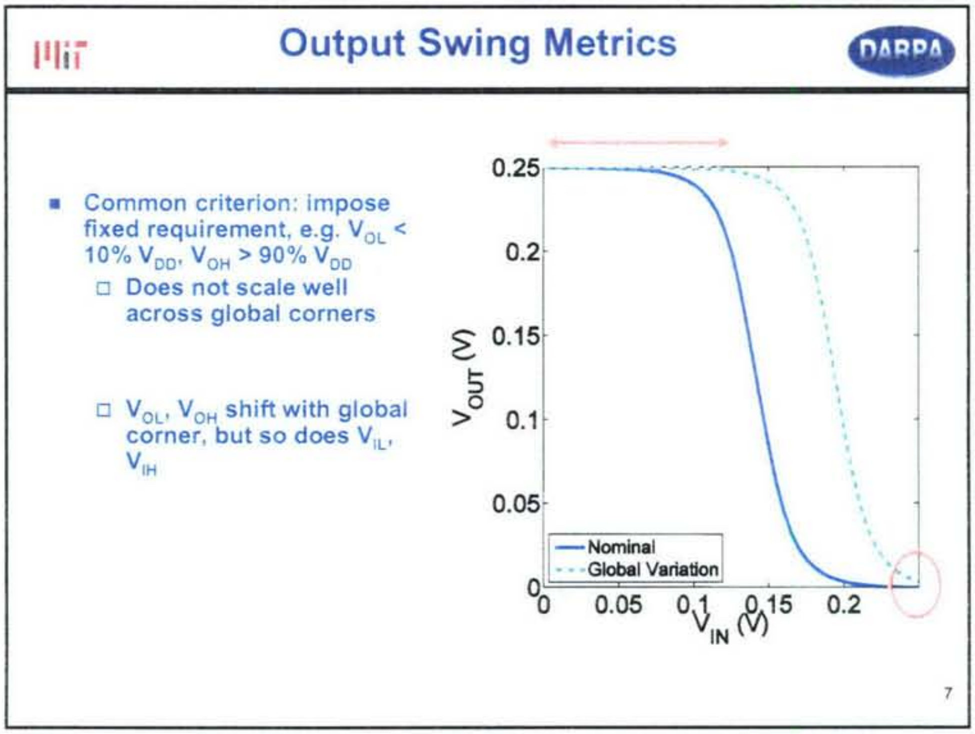
- V_T varies between transistors on the same die
  - □ Modeled as Gaussian distribution



Sub-V_T Logic Design Issues

- V_T variation causes distribution of 'high' and 'low' voltage levels of a logic gate

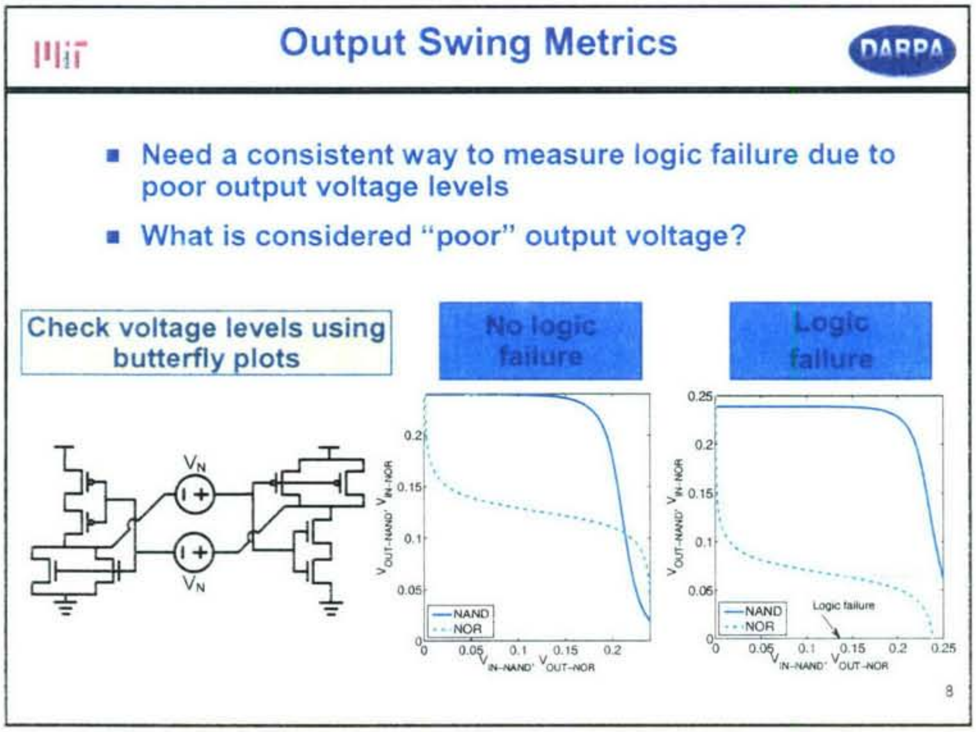
What V_{OL}, V_{OH} levels are acceptable?



Output Swing Metrics

- Common criterion: impose fixed requirement, e.g. V_{OL} < 10% V_{DD}, V_{OH} > 90% V_{DD}
  - □ Does not scale well across global corners
  - □ V_{OL}, V_{OH} shift with global corner, but so does V_{IL}, V_{IH}

0.25
0.2
0.15
0.1
0.05
0
V_OUT (V)
0
0.05
0.1
0.15
0.2
V_IN (V)
Nominal
Global Variation



Output Swing Metrics

- Need a consistent way to measure logic failure due to poor output voltage levels
- What is considered "poor" output voltage?

Check voltage levels using butterfly plots

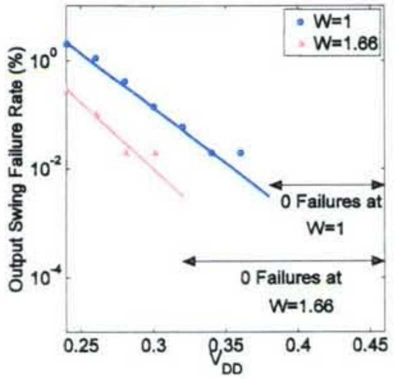
No logic failure

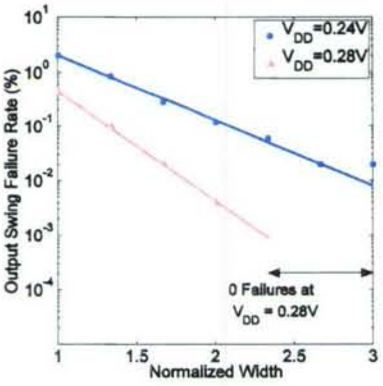
Logic failure

V_N
V_N
0.2
0.15
0.1
0.05
0
V_OUT-NAND, V_IN-NOR
0
0.05
0.1
0.15
0.2
V_IN-NAND, V_OUT-NOR
NAND
NOR
0.25
0.2
0.15
0.1
0.05
0
V_OUT-NAND, V_IN-NOR
0
0.05
0.1
0.15
0.2
0.25
V_IN-NAND, V_OUT-NOR
NAND
NOR
Logic failure

Output Swing Failure Rate

- Put logic gate under test back-to-back with:
  - □ NAND3 to check V_{OH}
  - □ NOR3 to check V_{OL}
- Define failure = no enclosed square in butterfly plots
- Failure rate decreases exponentially with device width and V_{DD}

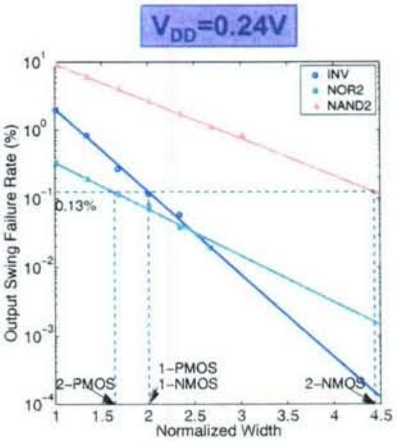


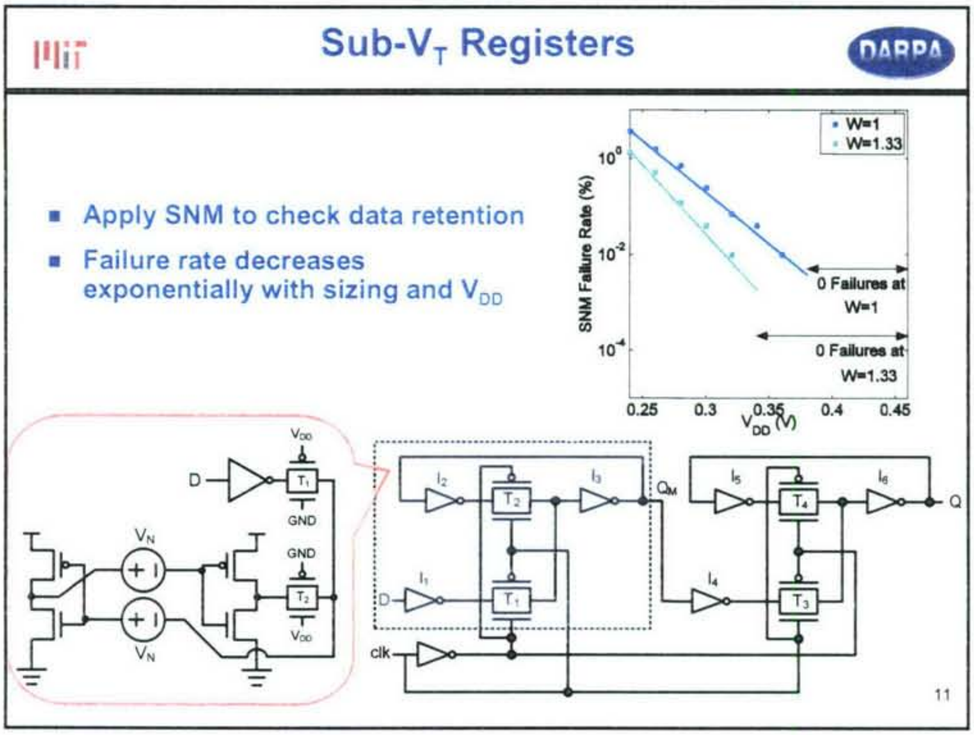


Constant-Yield Device Sizing

- Find failure rate vs. width plots for different circuit primitives
- Keep device sizes as small as possible, subject to yield constraint

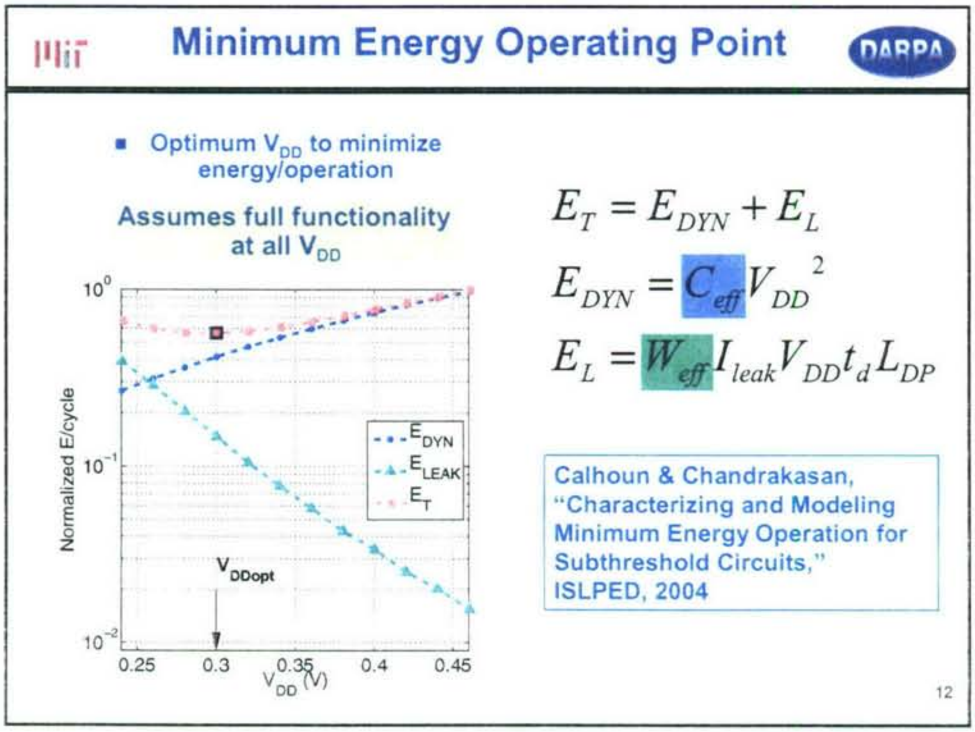
V_{DD}(V)	0.24	0.26	0.28	0.30	0.32	0.34
1-NMOS	2	1.67	1.33	1	1	1
1-PMOS	2	1.67	1.33	1	1	1
2-NMOS	4.43	2.93	2.3	2.27	1.3	1
2-PMOS	1.63	1	1	1	1	1





Sub-V_T Registers

- Apply SNM to check data retention
- Failure rate decreases exponentially with sizing and V_{DD}



Minimum Energy Operating Point

- Optimum V_{DD} to minimize energy/operation

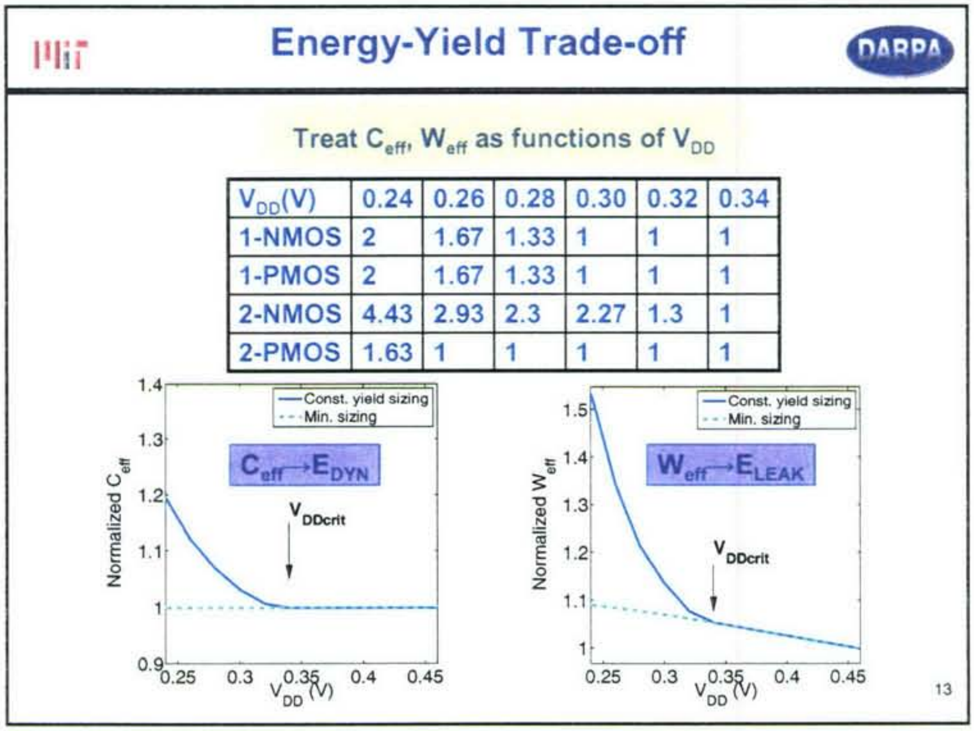
Assumes full functionality at all V_{DD}

$$E_T = E_{DYN} + E_L$$

$$E_{DYN} = C_{eff} V_{DD}{}^2$$

$$E_L = W_{eff} I_{leak} V_{DD} t_d L_{DP}$$

Calhoun & Chandrakasan, "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits," ISLPED, 2004



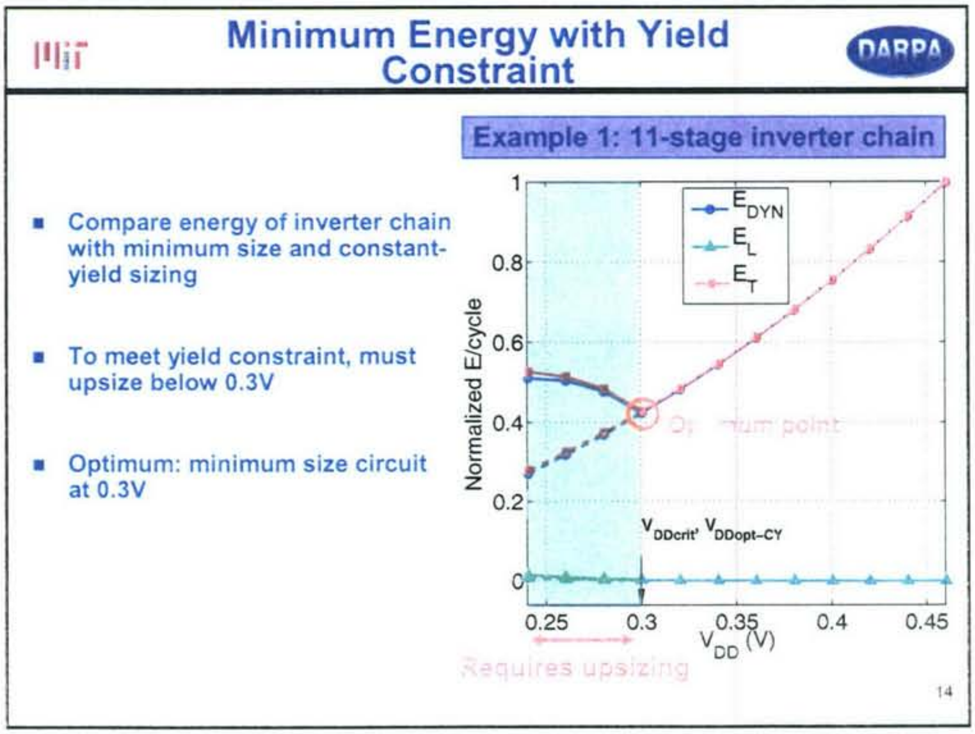
Energy-Yield Trade-off

Treat C_{eff}, W_{eff} as functions of V_{DD}

V_{DD}(V)	0.24	0.26	0.28	0.30	0.32	0.34
1-NMOS	2	1.67	1.33	1	1	1
1-PMOS	2	1.67	1.33	1	1	1
2-NMOS	4.43	2.93	2.3	2.27	1.3	1
2-PMOS	1.63	1	1	1	1	1

$C_{eff} \rightarrow E_{DYN}$

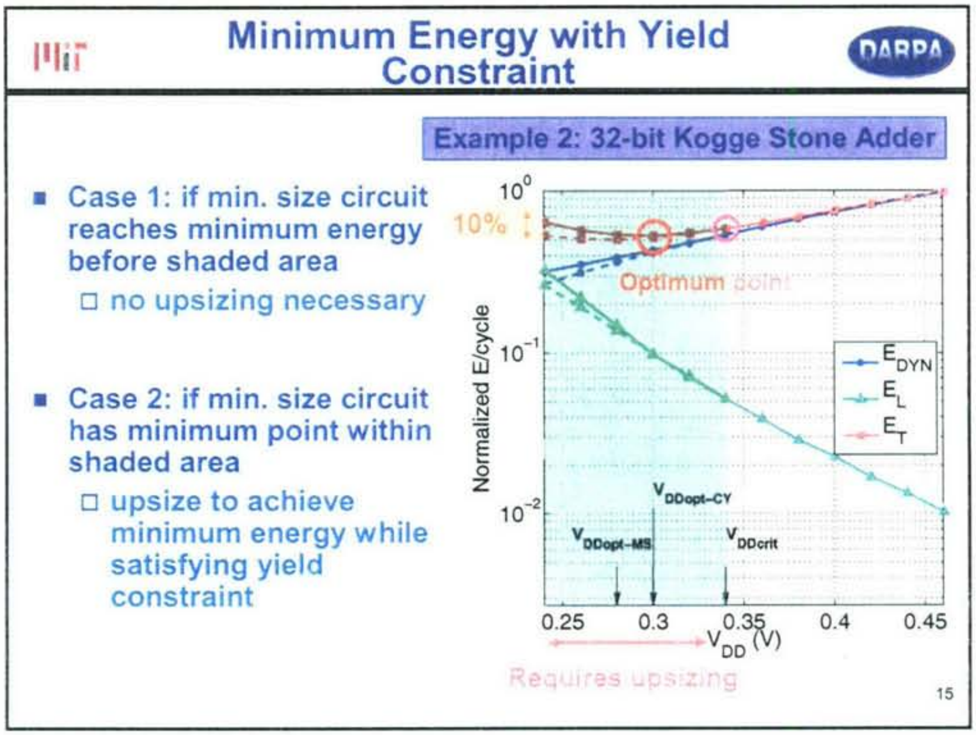
$W_{eff} \rightarrow E_{LEAK}$



Minimum Energy with Yield Constraint

- Compare energy of inverter chain with minimum size and constant-yield sizing
- To meet yield constraint, must upsize below 0.3V
- Optimum: minimum size circuit at 0.3V

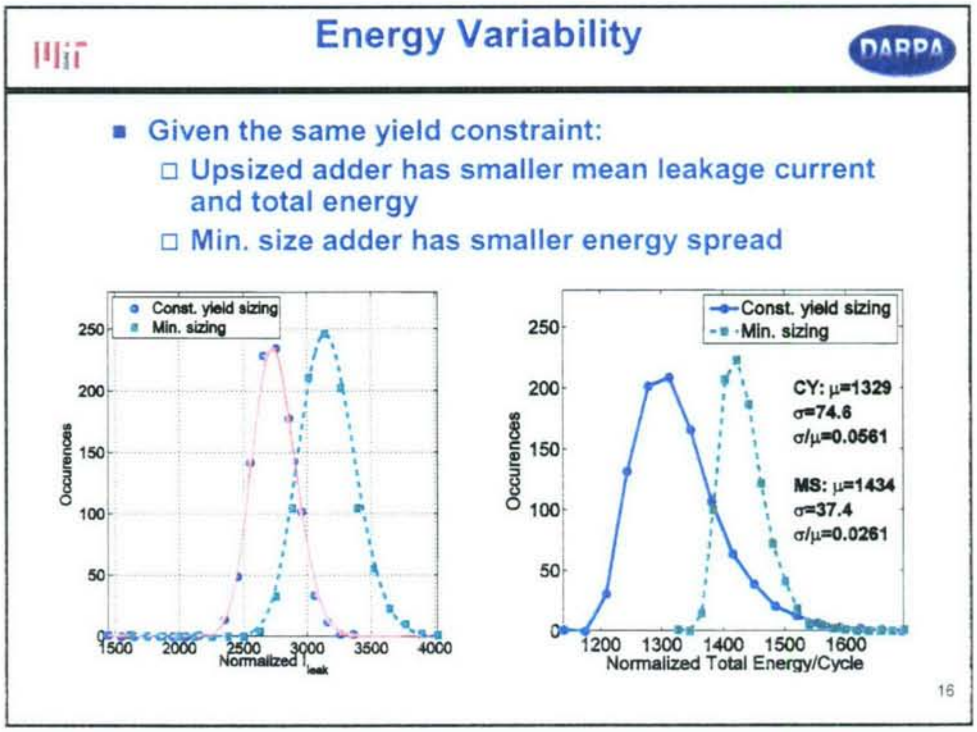
Example 1: 11-stage inverter chain



Minimum Energy with Yield Constraint

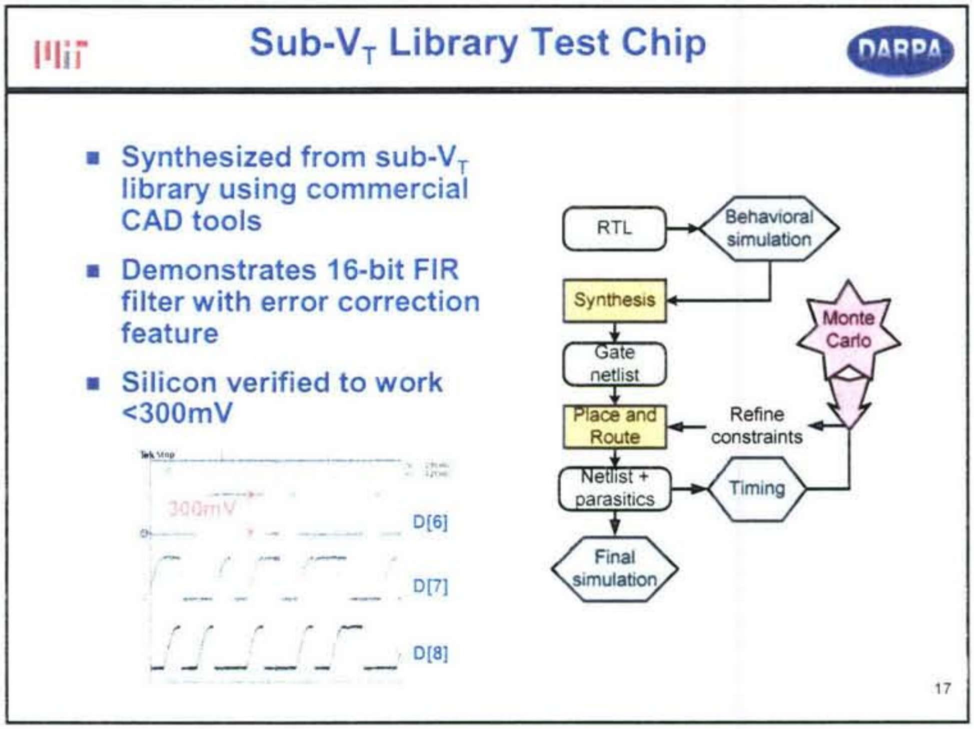
Example 2: 32-bit Kogge Stone Adder

- Case 1: if min. size circuit reaches minimum energy before shaded area
  - □ no upsizing necessary
- Case 2: if min. size circuit has minimum point within shaded area
  - □ upsize to achieve minimum energy while satisfying yield constraint



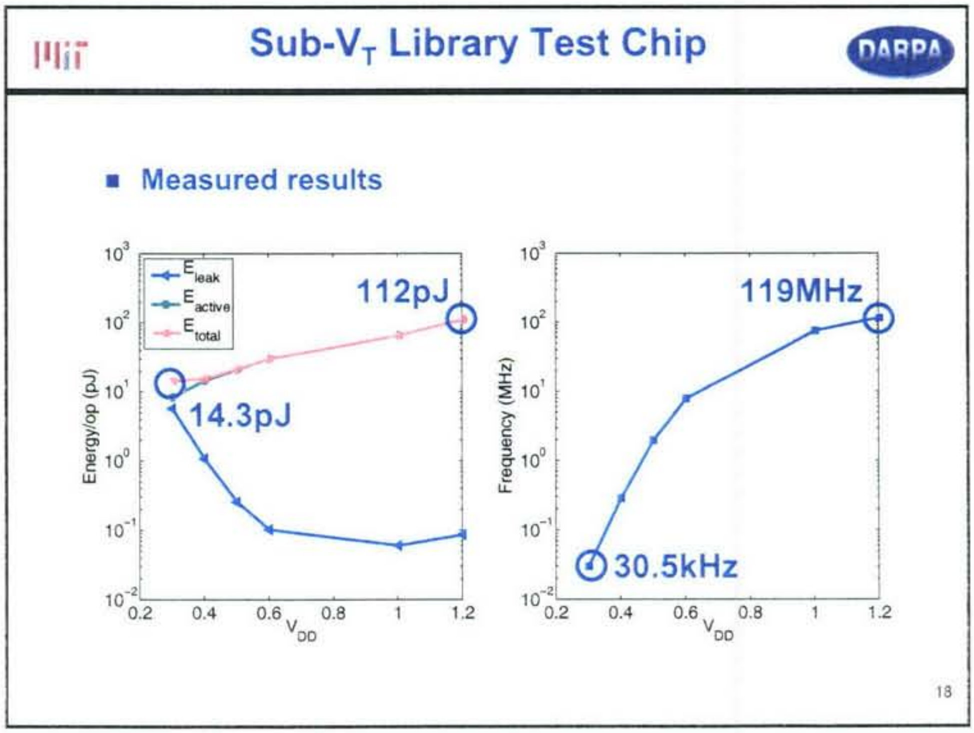
Energy Variability

- Given the same yield constraint:
  - □ Upsized adder has smaller mean leakage current and total energy
  - □ Min. size adder has smaller energy spread



Sub-V_T Library Test Chip

- Synthesized from sub-V_T library using commercial CAD tools
- Demonstrates 16-bit FIR filter with error correction feature
- Silicon verified to work <300mV

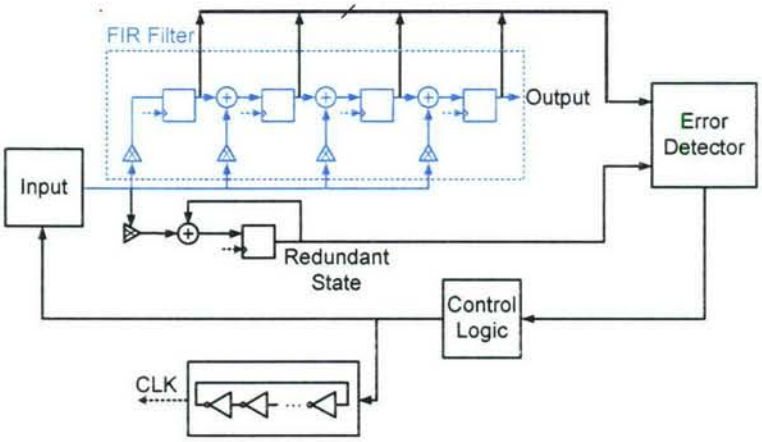


Sub-V_T Library Test Chip

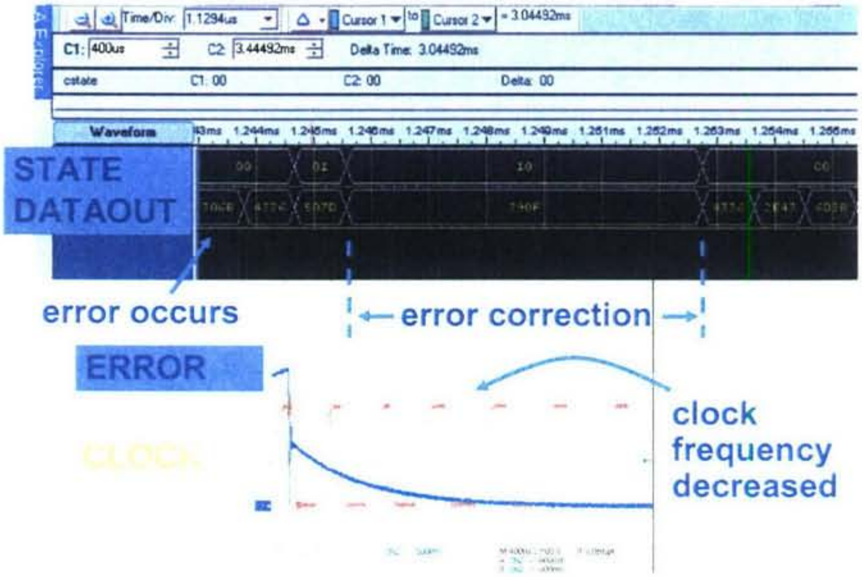
- Measured results

Error Correction

- Encode system states into redundant state
- Detect hard and soft errors
- Correct soft errors by adjusting clock frequency



Error Correction Demonstration



error occurs

← error correction →

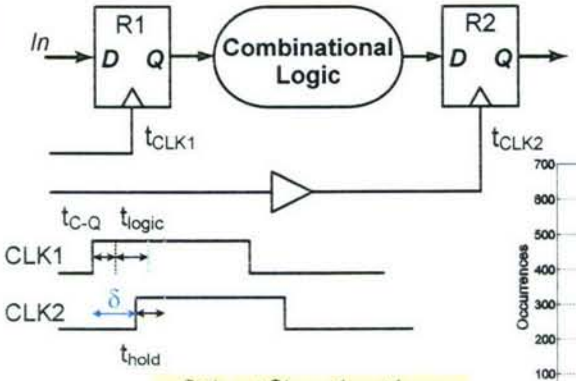
clock frequency decreased

Outline

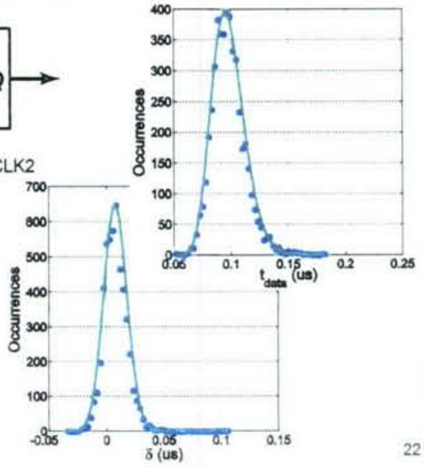
- Motivation
- Results
- Logic design issues in sub-V_T
- Minimum energy operation given yield constraint
- Sub-V_T library test chip
- Algorithmic error correction
- **Next phase**
  - □ Timing analysis
  - □ Integrated sub-V_T system

Timing Analysis

- Traditional STA uses deterministic best/worst case delay
- Leads to over-design when delay spread is large
- Consider statistical averaging effects

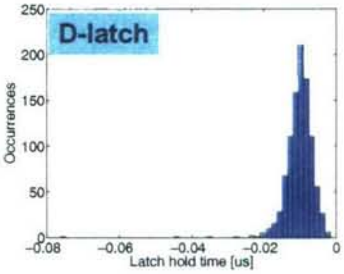


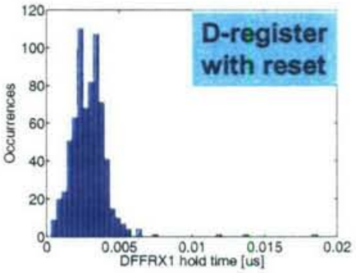
$\delta + t_{hold} + 2t_{jitter} < t_{C-Q} + t_{logic}$



Register Hold Time

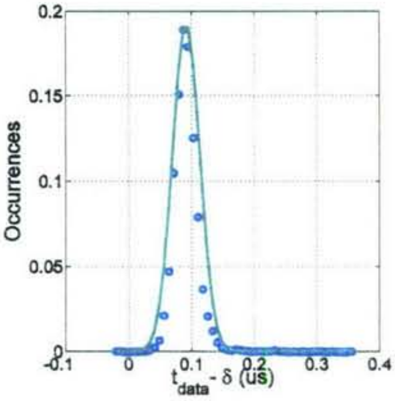
- Register hold time variability (σ/μ)
  - □ depends on clock/data slew rates
  - □ differs between registers
- No easy way to predict hold time of a given register!





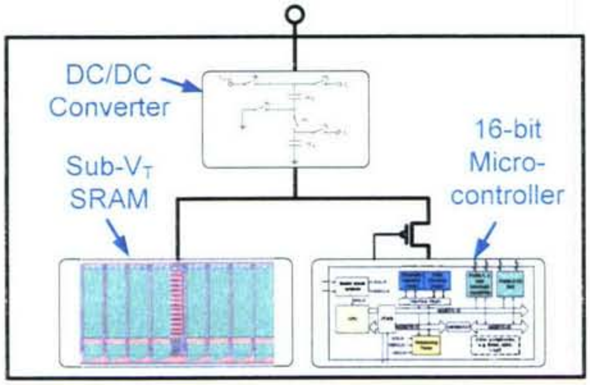
Possible Approaches

- Model as canonical distributions
  - □ makes computation easier
  - □ accurate modeling of distribution tail is critical
- Efficient simulation approach
  - □ simulate paths most likely to fail
  - □ coarsely discretize delay distribution



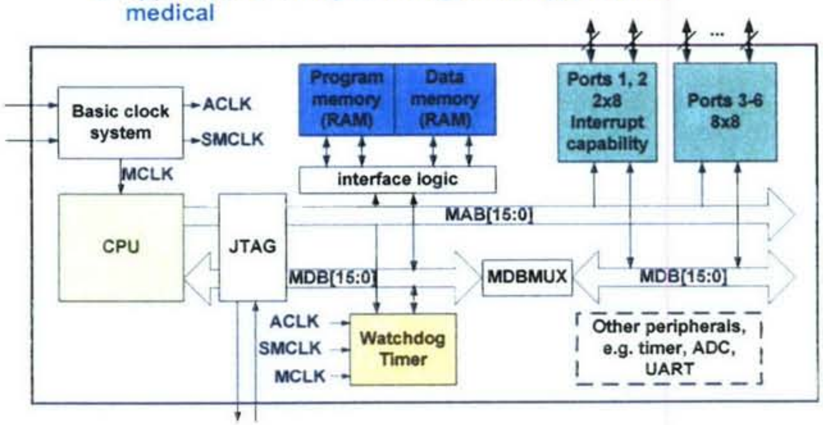
Integrated Sub-V_T System

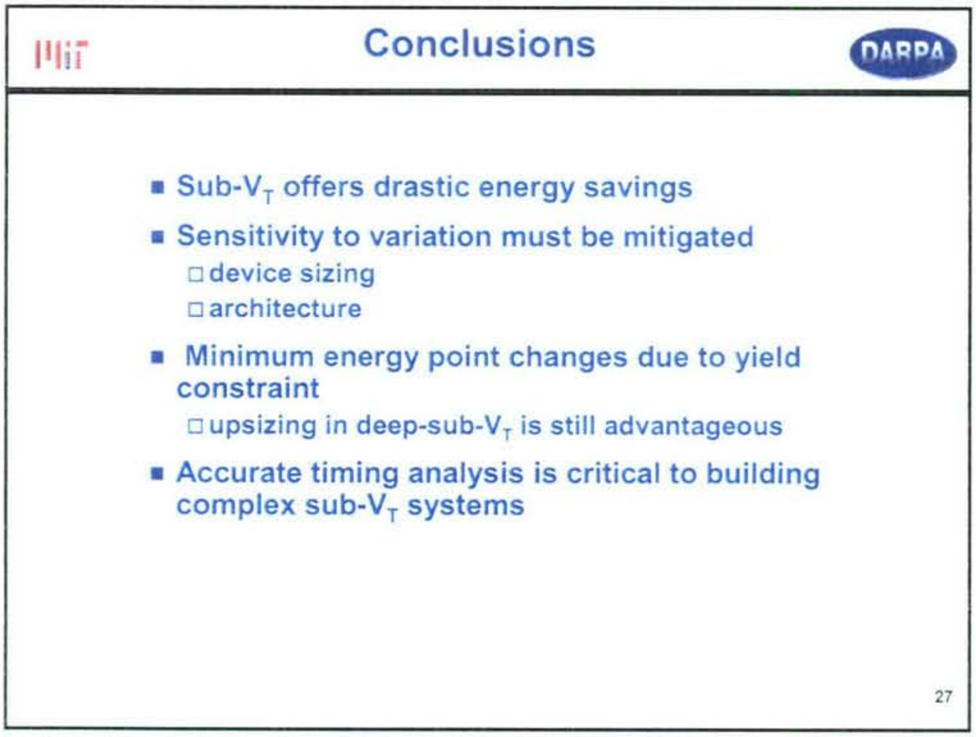
- Integrated system operating from external battery
- Sub-V_T SRAM critical to reducing leakage
- Microcontroller supports low power modes



Sub-V_T Microcontroller (MSP430)

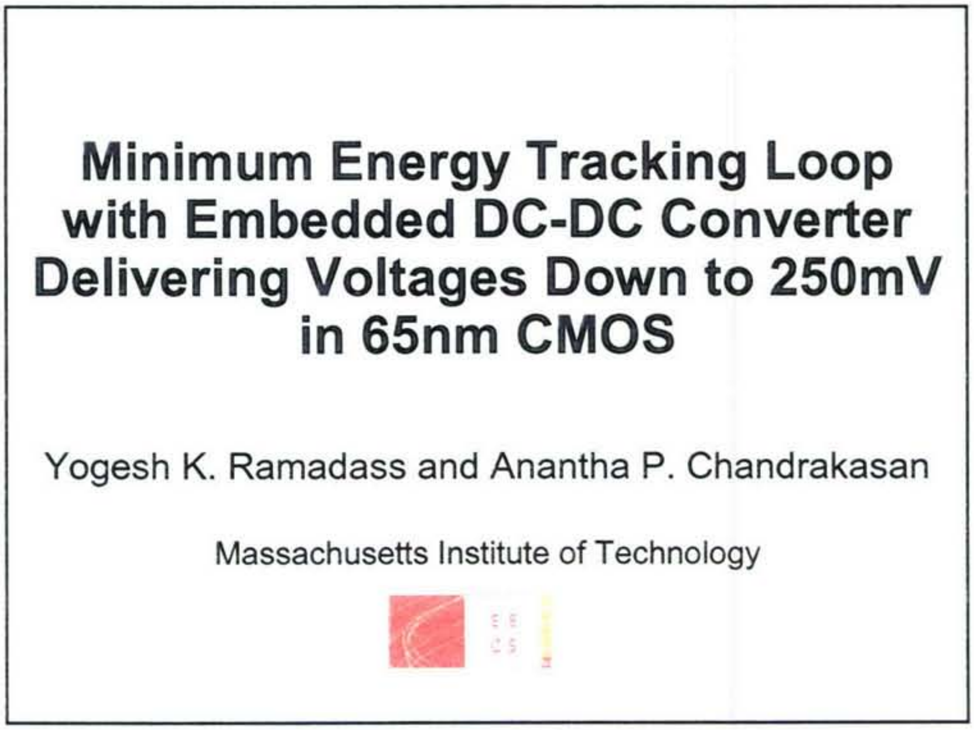
- TI MSP430
  - □ 16-bit RISC
  - □ 16 registers, 7 addressing modes, 27 instructions
  - □ applications: utility metering, security, portable medical





Conclusions

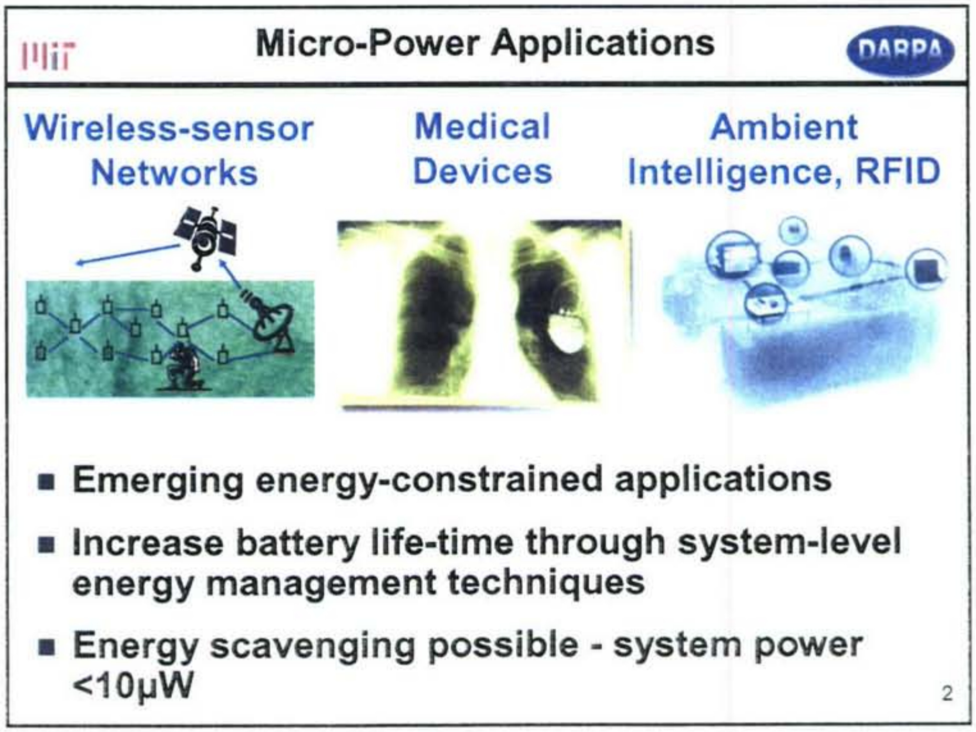
- Sub-V_T offers drastic energy savings
- Sensitivity to variation must be mitigated
  - □ device sizing
  - □ architecture
- Minimum energy point changes due to yield constraint
  - □ upsizing in deep-sub-V_T is still advantageous
- Accurate timing analysis is critical to building complex sub-V_T systems



Minimum Energy Tracking Loop with Embedded DC-DC Converter Delivering Voltages Down to 250mV in 65nm CMOS

Yogesh K. Ramadass and Anantha P. Chandrakasan

Massachusetts Institute of Technology



Micro-Power Applications

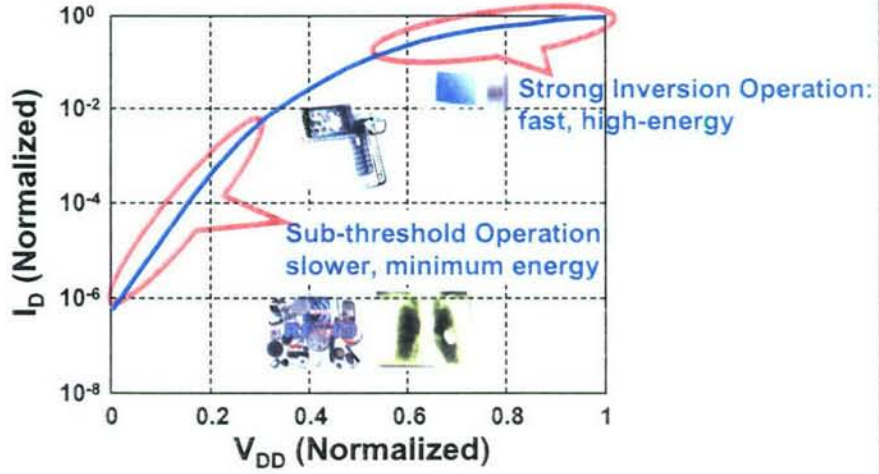
Wireless-sensor Networks	Medical Devices	Ambient Intelligence, RFID

- Emerging energy-constrained applications
- Increase battery life-time through system-level energy management techniques
- Energy scavenging possible - system power <10μW

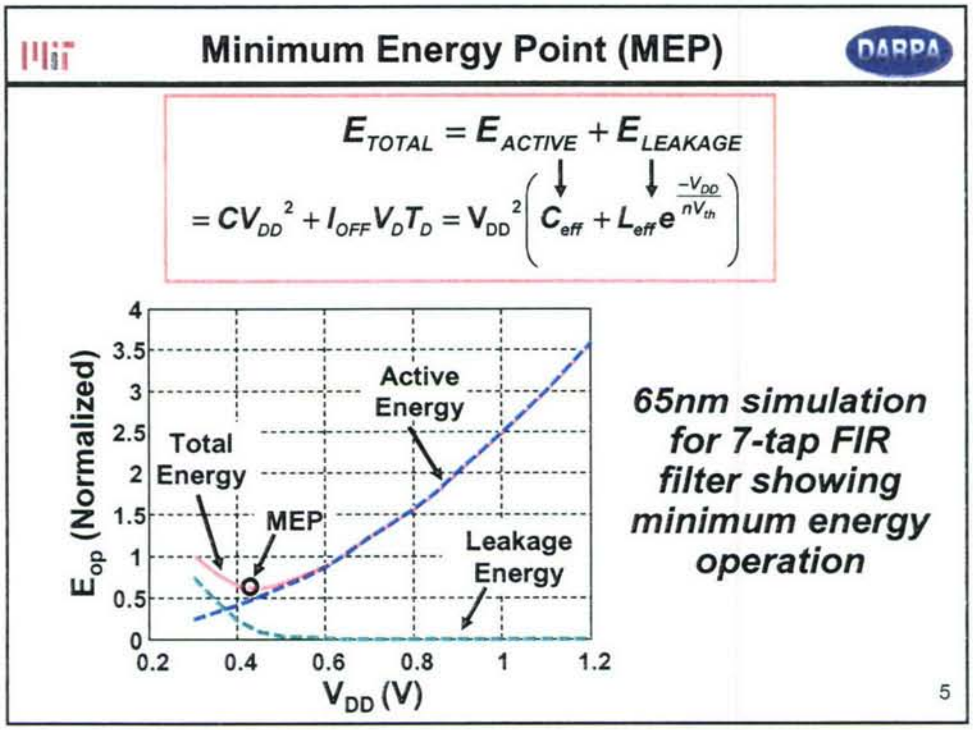
Outline

- Motivation and System Architecture
- Energy Sensing Technique
- Low Power DC-DC Converter
- Measurement Results
- Conclusions

Sub-threshold Operation



Goal: Minimize Energy per Operation



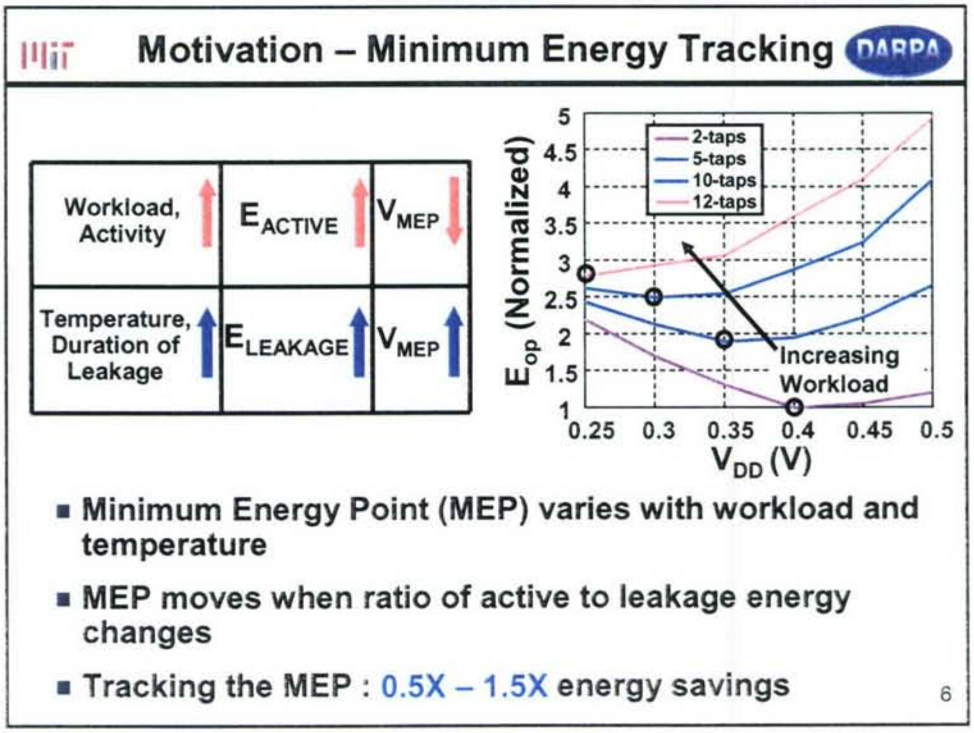
Minimum Energy Point (MEP)

$$E_{TOTAL} = E_{ACTIVE} + E_{LEAKAGE}$$

$$\downarrow \qquad \downarrow$$

$$= CV_{DD}^{2} + I_{OFF}V_{D}T_{D} = \mathrm{V_{DD}}^{2}\left(C_{eff} + L_{eff}e^{\frac{-V_{DD}}{nV_{th}}}\right)$$

65nm simulation for 7-tap FIR filter showing minimum energy operation

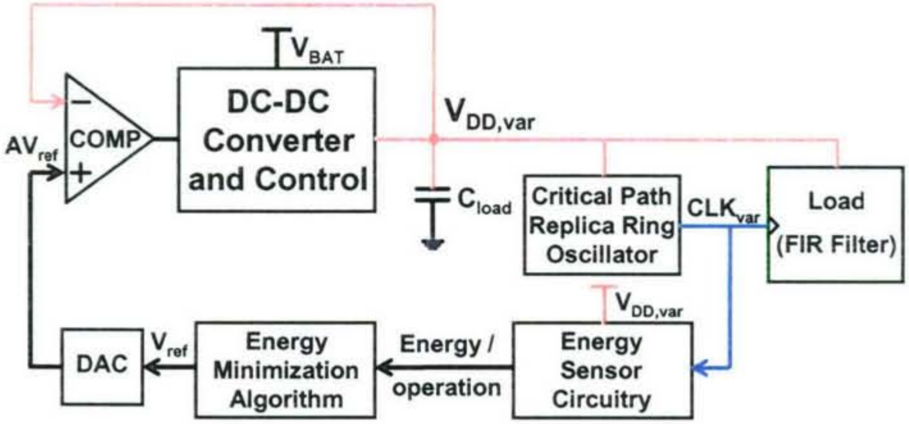


Motivation – Minimum Energy Tracking

Workload, Activity ↑	E_{ACTIVE} ↑	V_{MEP} ↓
Temperature, Duration of Leakage ↑	$E_{LEAKAGE}$ ↑	V_{MEP} ↑

- Minimum Energy Point (MEP) varies with workload and temperature
- MEP moves when ratio of active to leakage energy changes
- Tracking the MEP : 0.5X – 1.5X energy savings

Minimum Energy Tracking Loop



- Completely on-chip except for the passive filter components

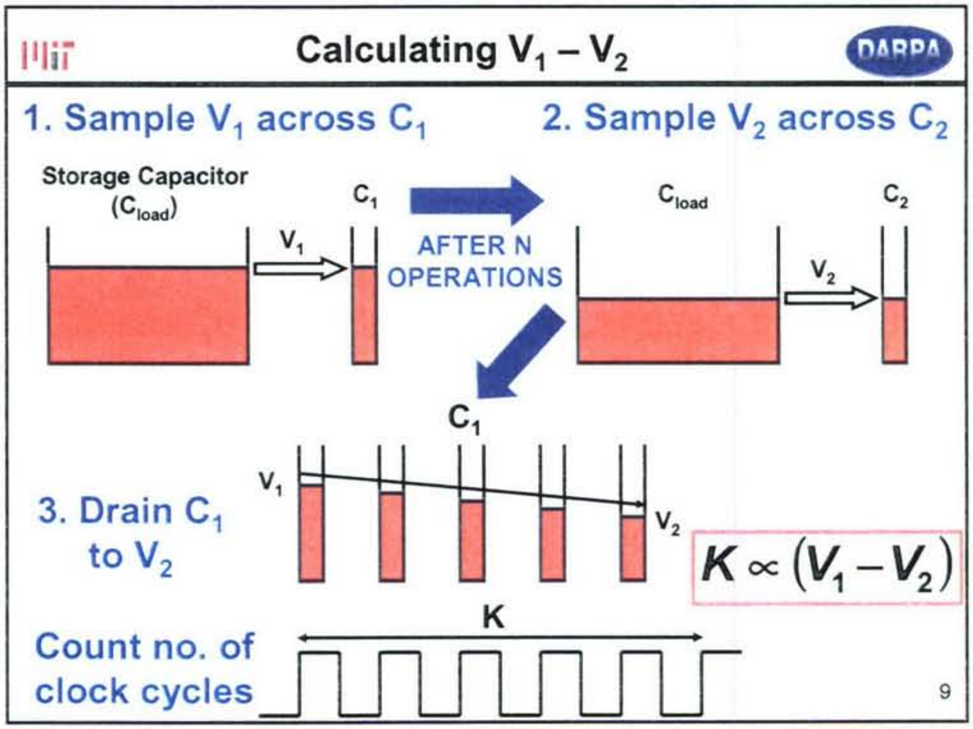
Energy/Operation (E_{op}) Formulation

- Voltage across storage capacitor C_{load} droops from V_1 to V_2 in the course of N (i.e. 32 or 64) operations.

$$E_{op} = \frac{C_{load}}{2\times N}\left(V_1^2 - V_2^2\right) = \frac{C_{load}}{2\times N}\left(V_1 + V_2\right)\left(V_1 - V_2\right) \approx \frac{C_{load}}{N} V_1\left(V_1 - V_2\right)$$

$$\boxed{E_{op} \propto V_1\left(V_1 - V_2\right)}$$

- V_1 is set digitally as V_{ref} to the DC-DC converter
- $(V_1 - V_2)$ needs to be estimated



Calculating $V_1 - V_2$

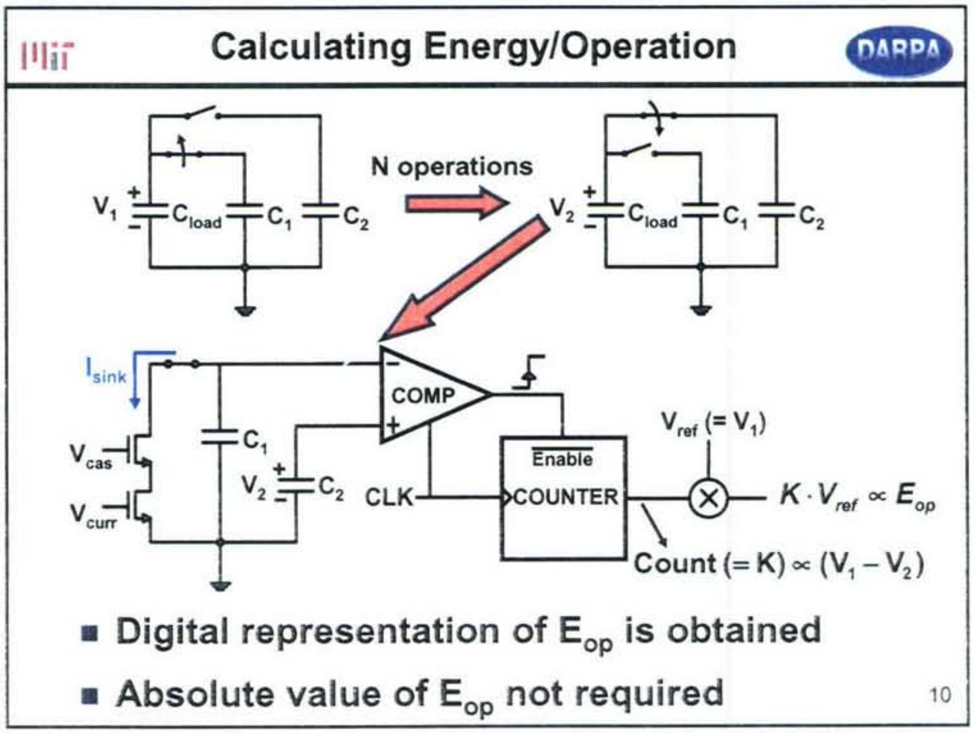
1. Sample V_1 across C_1

2. Sample V_2 across C_2

3. Drain C_1 to V_2

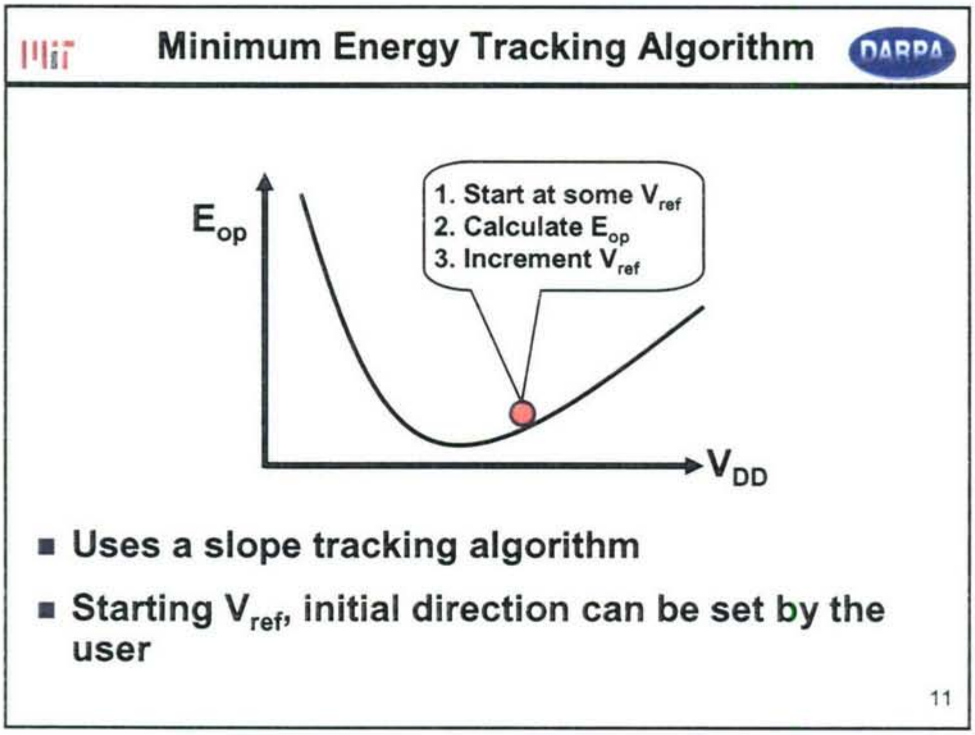
Count no. of clock cycles

$$K \propto (V_1 - V_2)$$



Calculating Energy/Operation

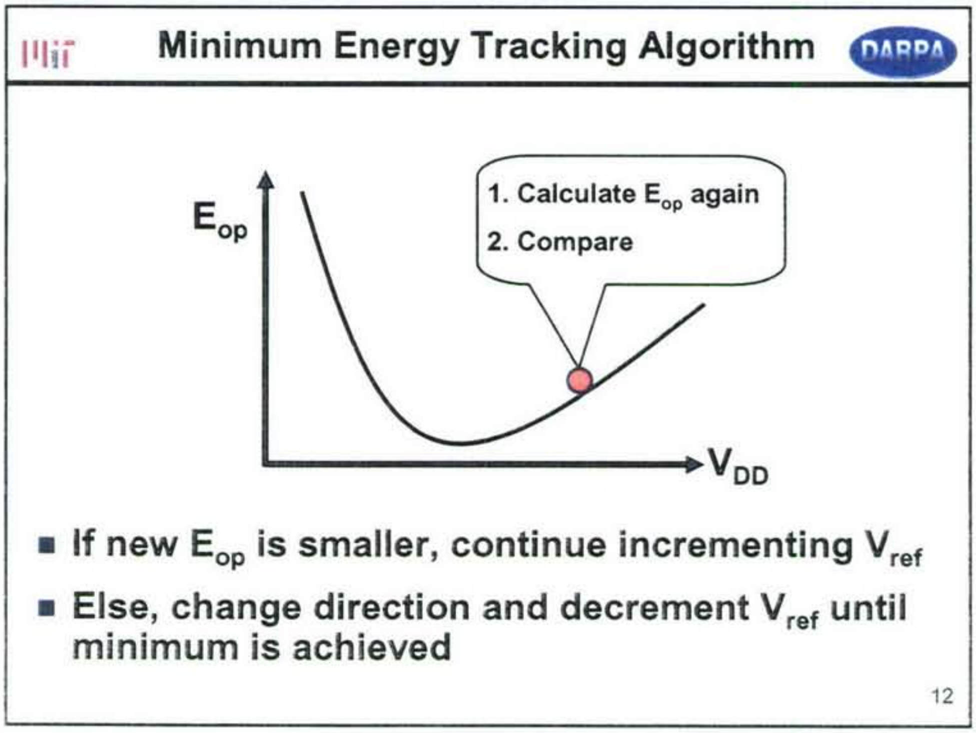
- Digital representation of E_{op} is obtained
- Absolute value of E_{op} not required



Minimum Energy Tracking Algorithm

1. Start at some V_{ref}
2. Calculate E_{op}
3. Increment V_{ref}

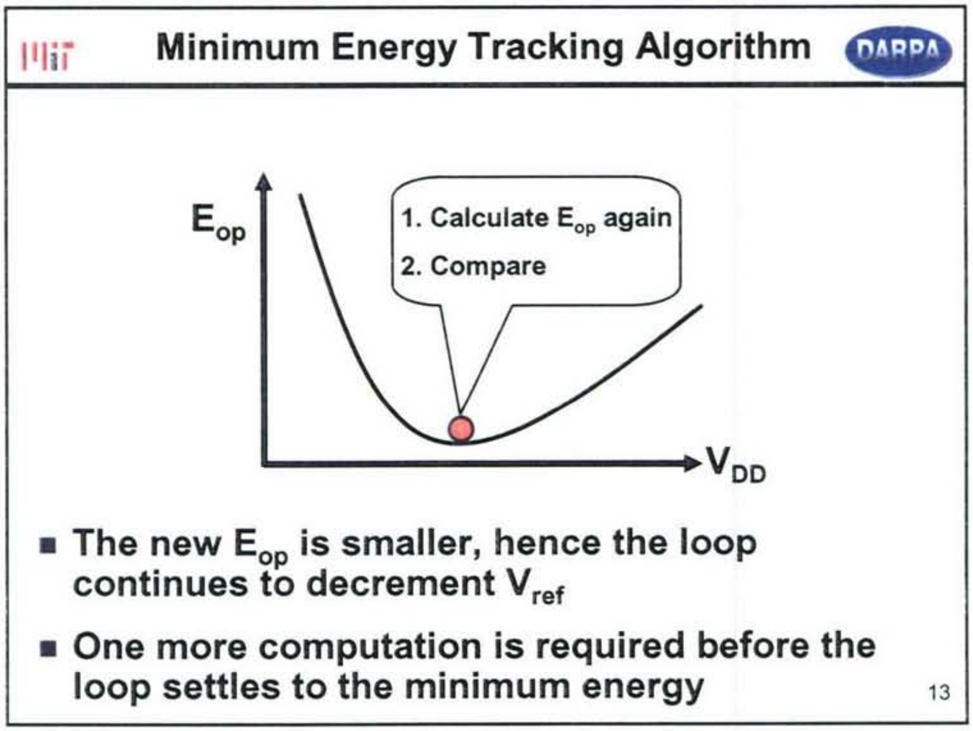
- Uses a slope tracking algorithm
- Starting V_{ref}, initial direction can be set by the user



Minimum Energy Tracking Algorithm

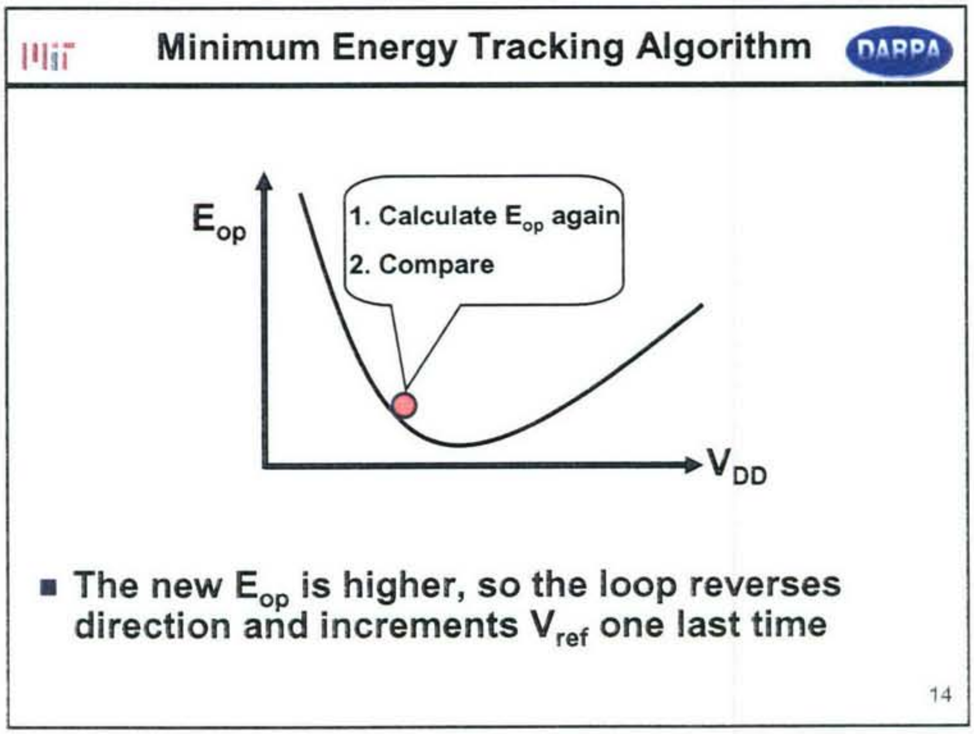
1. Calculate E_{op} again
2. Compare

- If new E_{op} is smaller, continue incrementing V_{ref}
- Else, change direction and decrement V_{ref} until minimum is achieved



Minimum Energy Tracking Algorithm

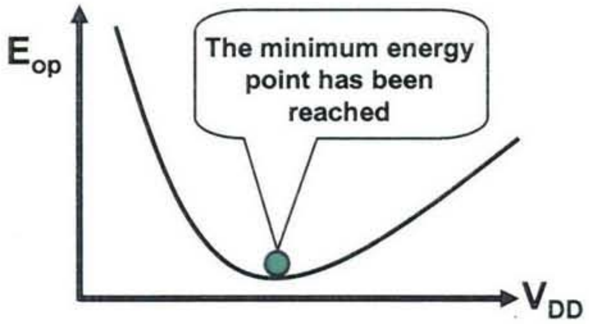
- The new E_{op} is smaller, hence the loop continues to decrement V_{ref}
- One more computation is required before the loop settles to the minimum energy



Minimum Energy Tracking Algorithm

- The new E_{op} is higher, so the loop reverses direction and increments V_{ref} one last time

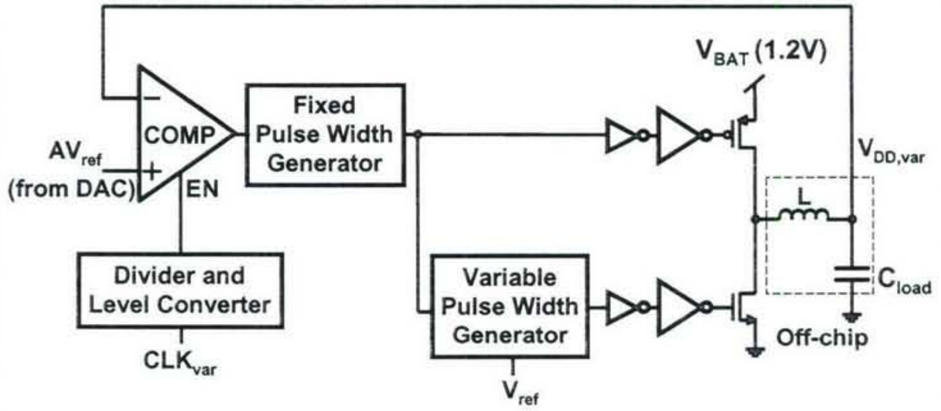
Minimum Energy Tracking Algorithm



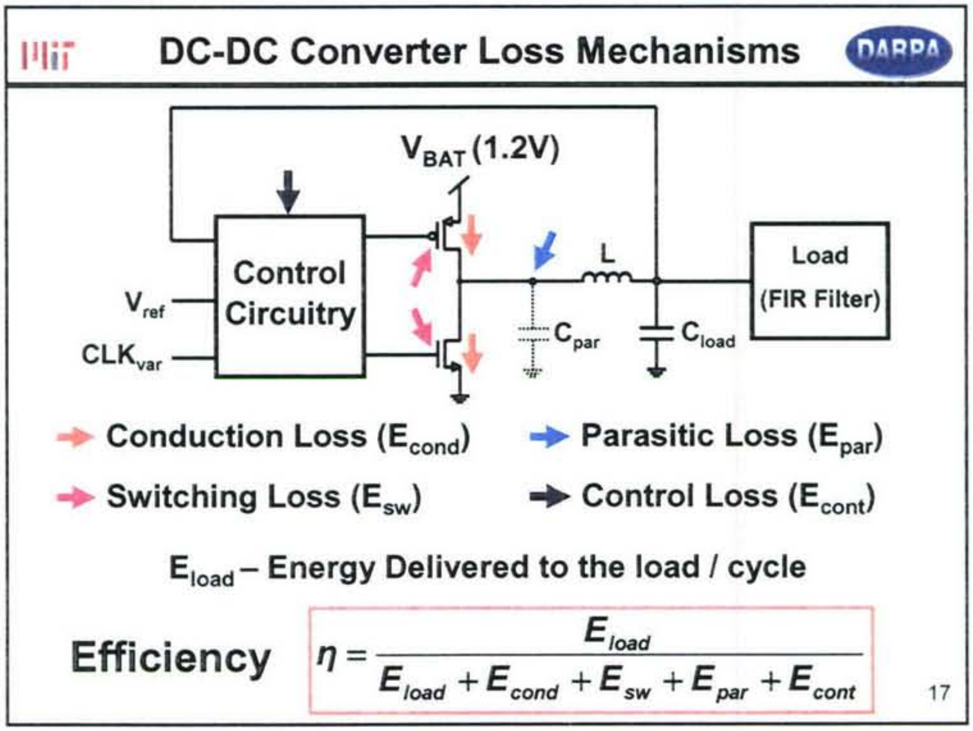
The minimum energy point has been reached

- V_{ref} is set at the minimum energy point
- The loop shuts down
- The load circuit continues to operate

DC-DC Converter Architecture



- V_{DD} : 250mV – 700mV ; Load Power : 1μW – 100μW
- Converter operates in Pulse Frequency Modulation (PFM) mode
- V_{ref} is set digitally by the loop



DC-DC Converter Loss Mechanisms

🟧 Conduction Loss (E_{cond})

🟥 Switching Loss (E_{sw})

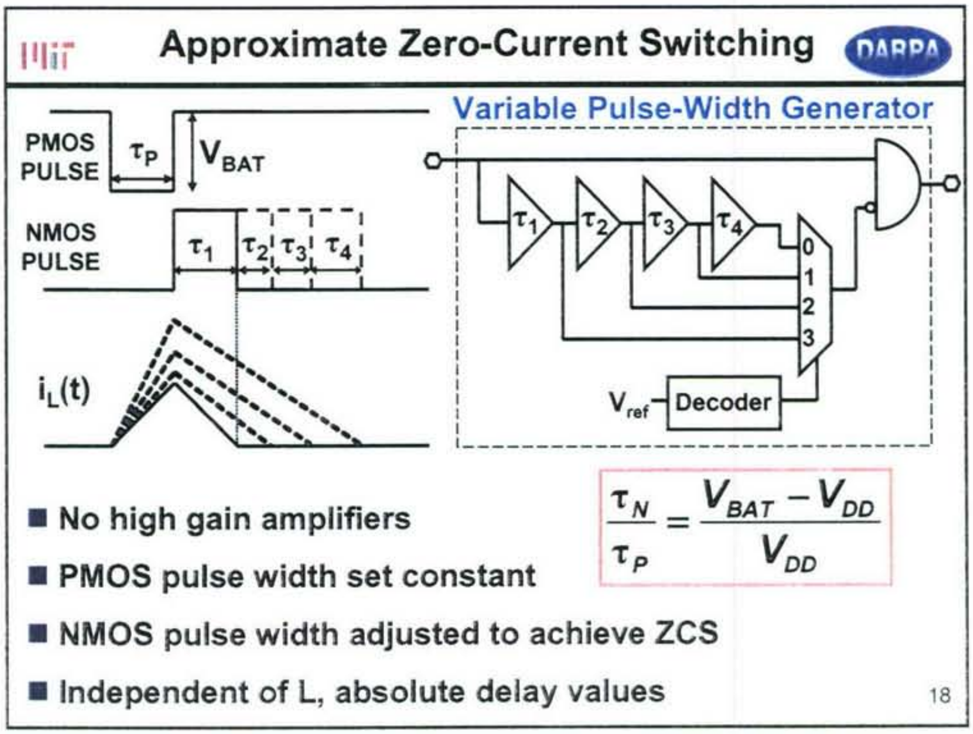
🟦 Parasitic Loss (E_{par})

⬛ Control Loss (E_{cont})

E_{load} – Energy Delivered to the load / cycle

Efficiency

$$\eta = \frac{E_{load}}{E_{load} + E_{cond} + E_{sw} + E_{par} + E_{cont}}$$



Approximate Zero-Current Switching

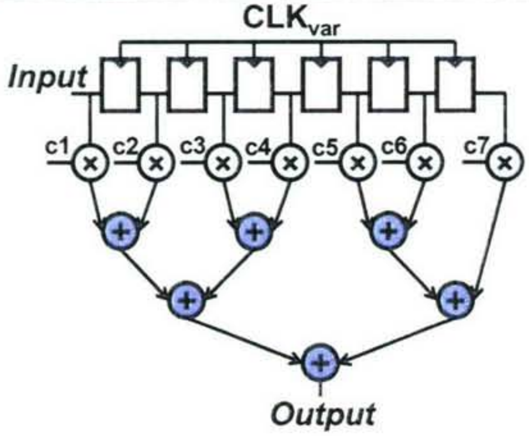
- No high gain amplifiers
- PMOS pulse width set constant
- NMOS pulse width adjusted to achieve ZCS
- Independent of L, absolute delay values

$$\frac{\tau_N}{\tau_P} = \frac{V_{BAT} - V_{DD}}{V_{DD}}$$

Techniques to Improve Efficiency

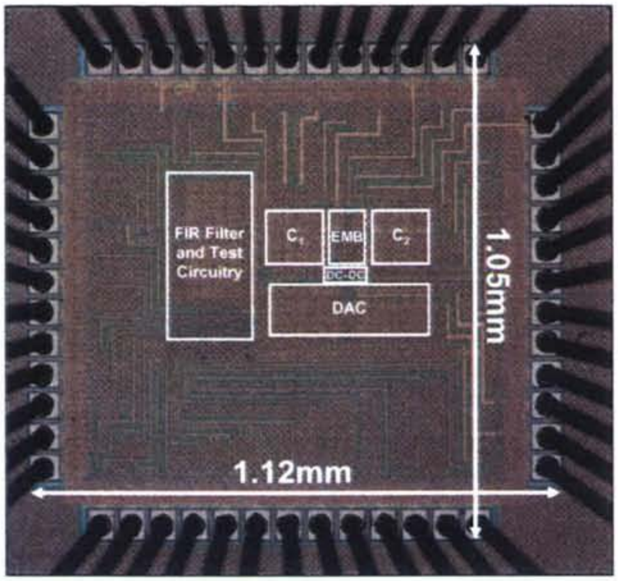
- Conduction, Switching Loss
  - ❑ Optimal Power Transistor Sizing
- Control Loss (affects low load efficiency)
  - ❑ Simple PFM Control
  - ❑ All-digital control to achieve approximate ZCS
  - ❑ Comparator clock scales with load power
- Parasitic Capacitance Loss
  - ❑ Finite delay between PMOS and NMOS pulses
  - ❑ Increase E_{load}, minimize contribution of E_{par}

Test Load - FIR Filter



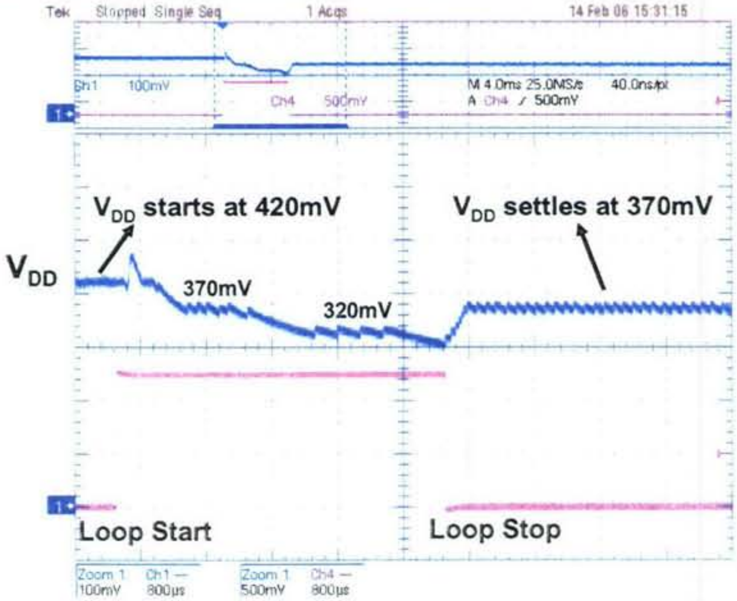
- 7-tap FIR filter capable of operation down to 250mV
- Workload varied by changing the number of taps
- Leakage remains constant as number of taps are changed

Test Chip Die Photo



- 65nm,6LM CMOS
- Die area – 1.05mm x 1.12mm
- Circuit active area – 0.23 mm²
- Minimum energy tracking circuitry occupies just 0.05mm²

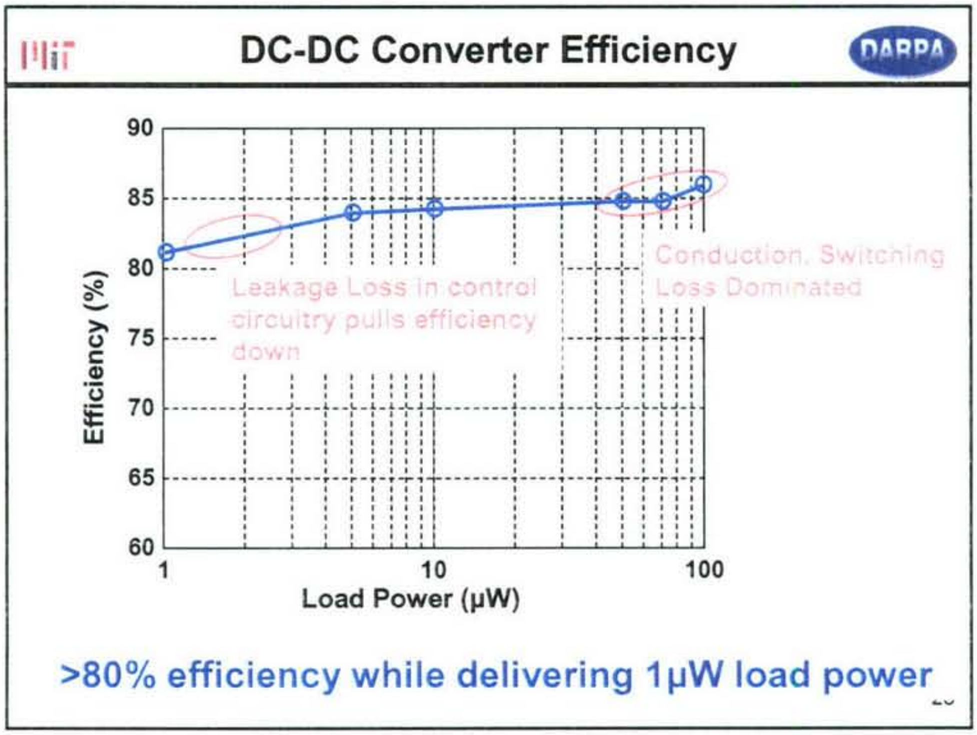
Operation of the Energy Minimizing Loop



1-tap

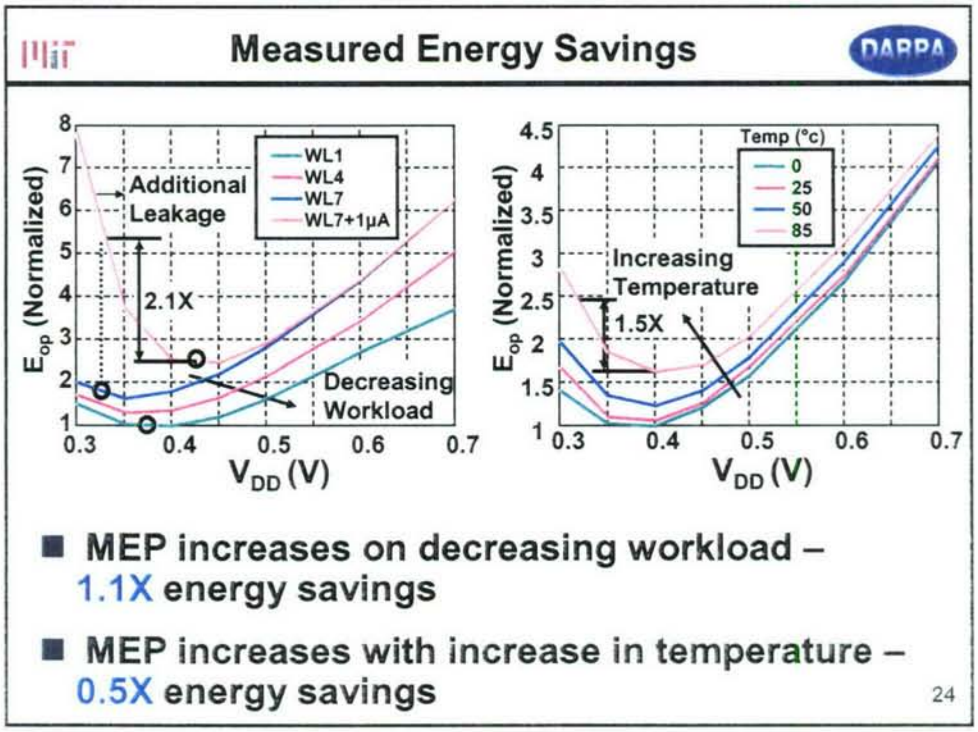
MEP

= 370mV



DC-DC Converter Efficiency

>80% efficiency while delivering 1µW load power



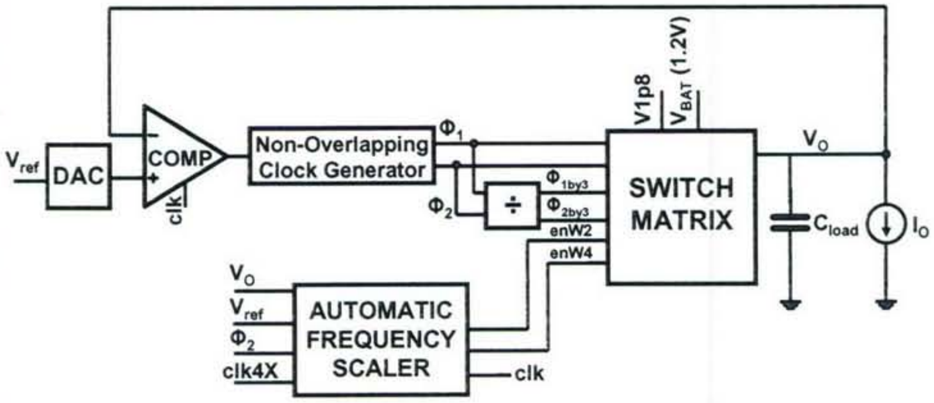
Measured Energy Savings

- MEP increases on decreasing workload – 1.1X energy savings
- MEP increases with increase in temperature – 0.5X energy savings

MEP Tracking Loop Properties

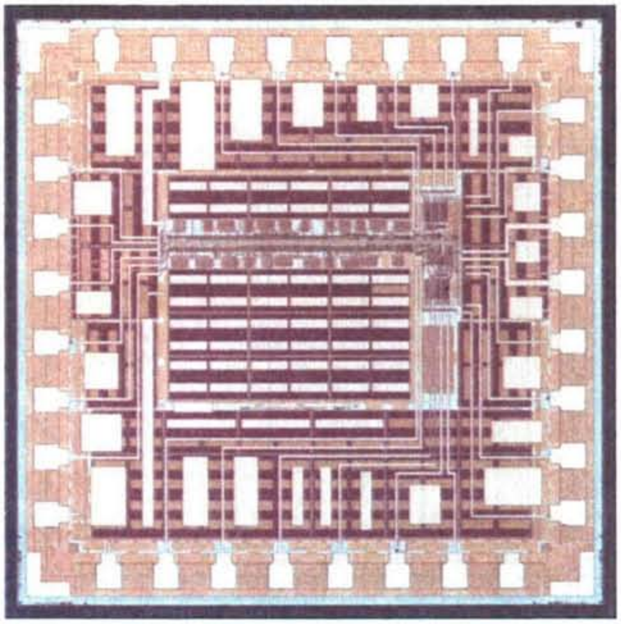
- Tracking Loop
  - □ Non-invasive tracking
  - □ Energy computed of the actual circuit – no replicas
- Tracking Methodology
  - □ Independent of the size of the load circuit
  - □ Independent of the DC-DC converter topology
- Overhead
  - □ Energy overhead ~ 50 operations
  - □ Area overhead = 0.05mm^2
  - □ Multiple loops – distinct voltage domains

Switched Cap Architecture



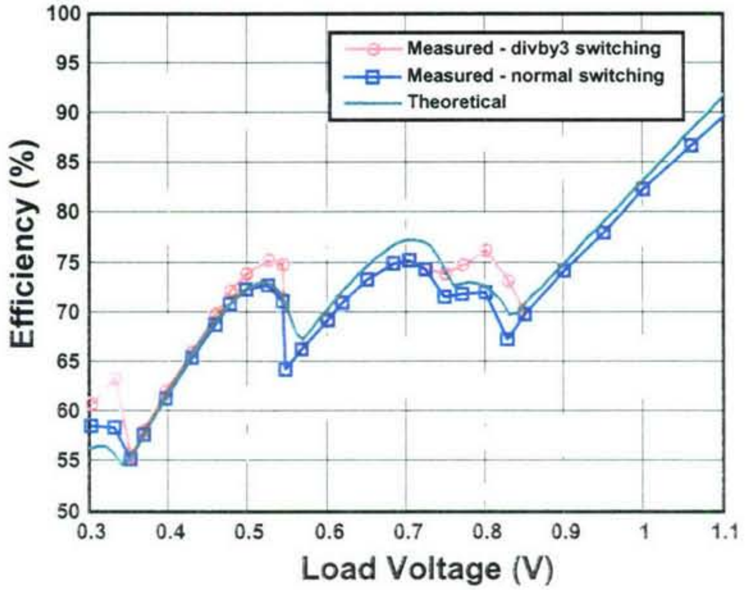
- ON-OFF mode control. V_{ref} is set digitally
- No static power loss
- Completely on-chip except for C_{load}

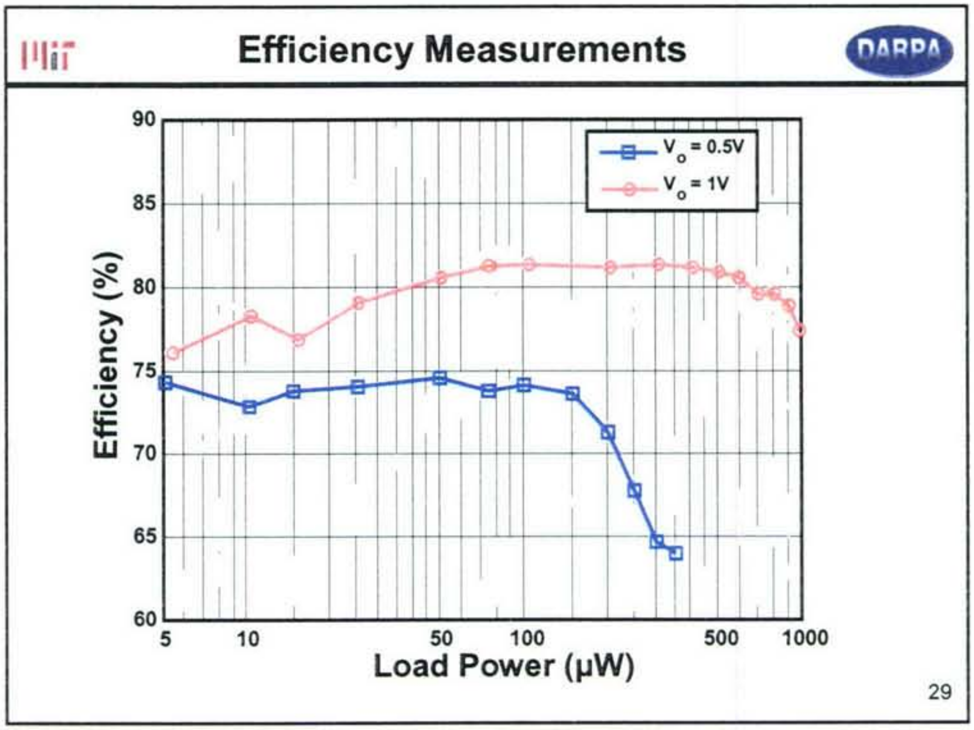
Die Photo



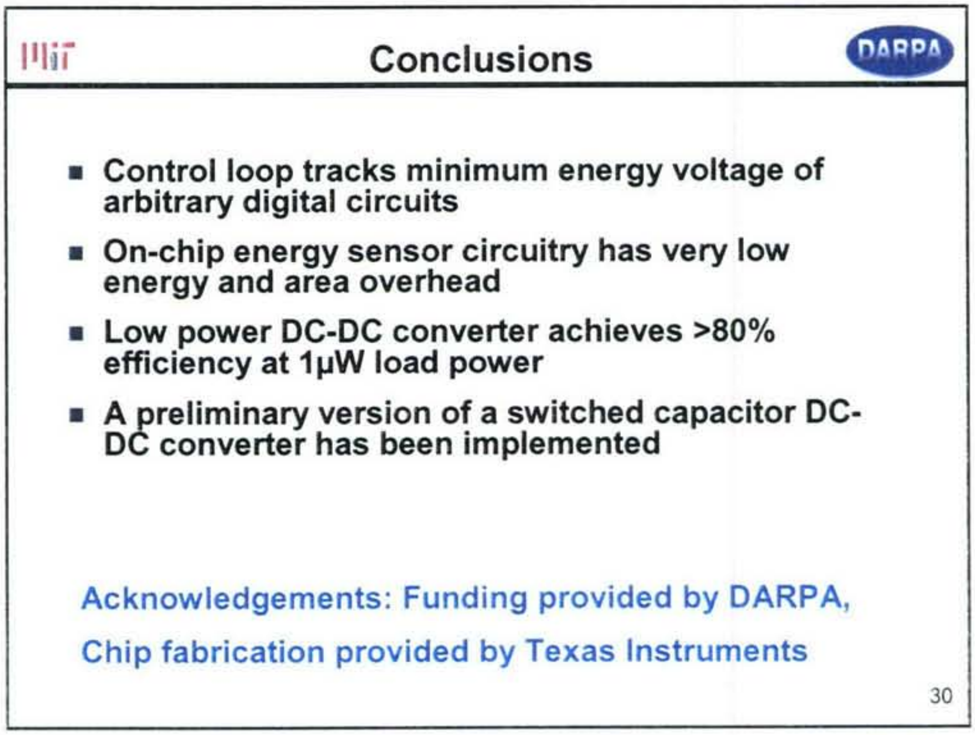
- Die area – 1.6mm x 1.6mm
- Circuit active area – 0.57 mm^2
- Gate-oxide capacitors are used

Efficiency Measurements





Efficiency Measurements

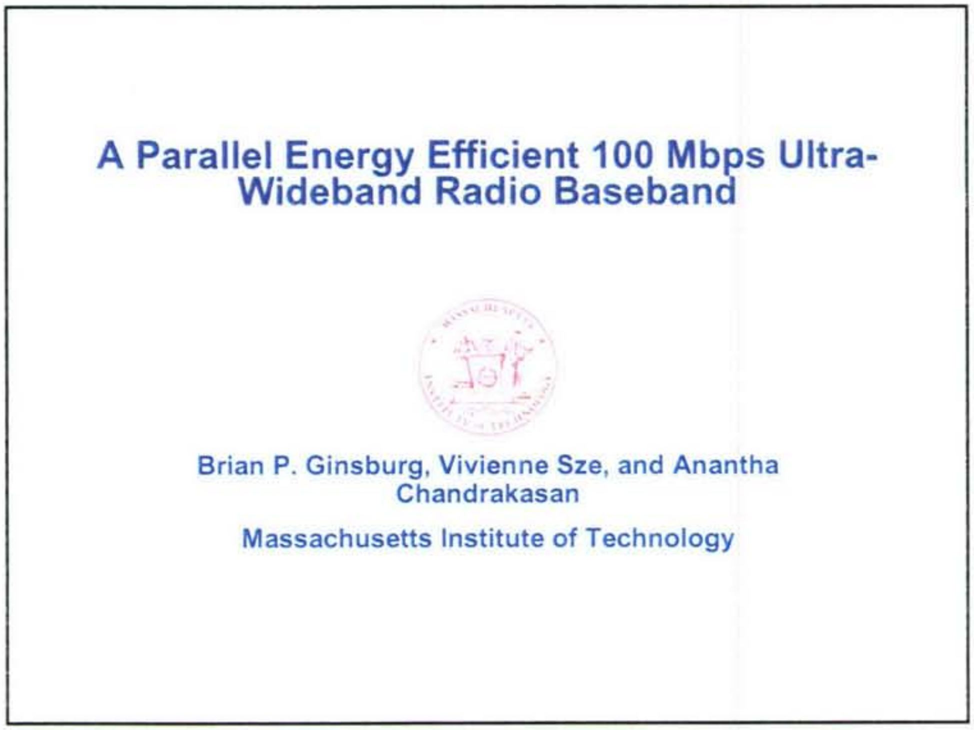


Conclusions

- Control loop tracks minimum energy voltage of arbitrary digital circuits
- On-chip energy sensor circuitry has very low energy and area overhead
- Low power DC-DC converter achieves >80% efficiency at 1µW load power
- A preliminary version of a switched capacitor DC-DC converter has been implemented

Acknowledgements: Funding provided by DARPA,

Chip fabrication provided by Texas Instruments



A Parallel Energy Efficient 100 Mbps Ultra-Wideband Radio Baseband

Brian P. Ginsburg, Vivienne Sze, and Anantha Chandrakasan

Massachusetts Institute of Technology

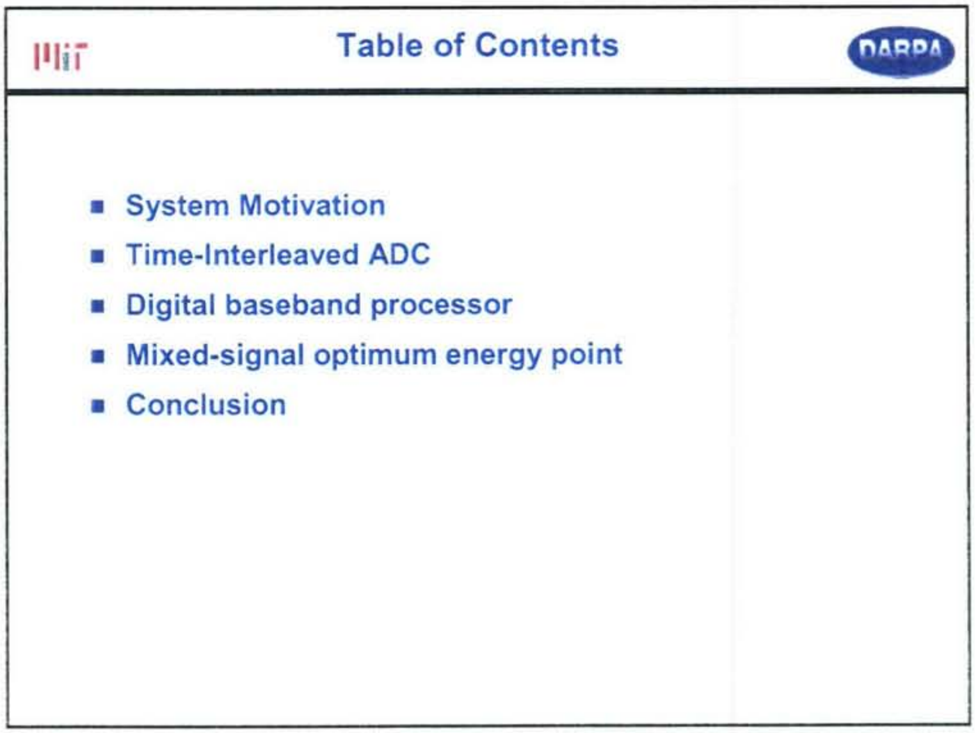
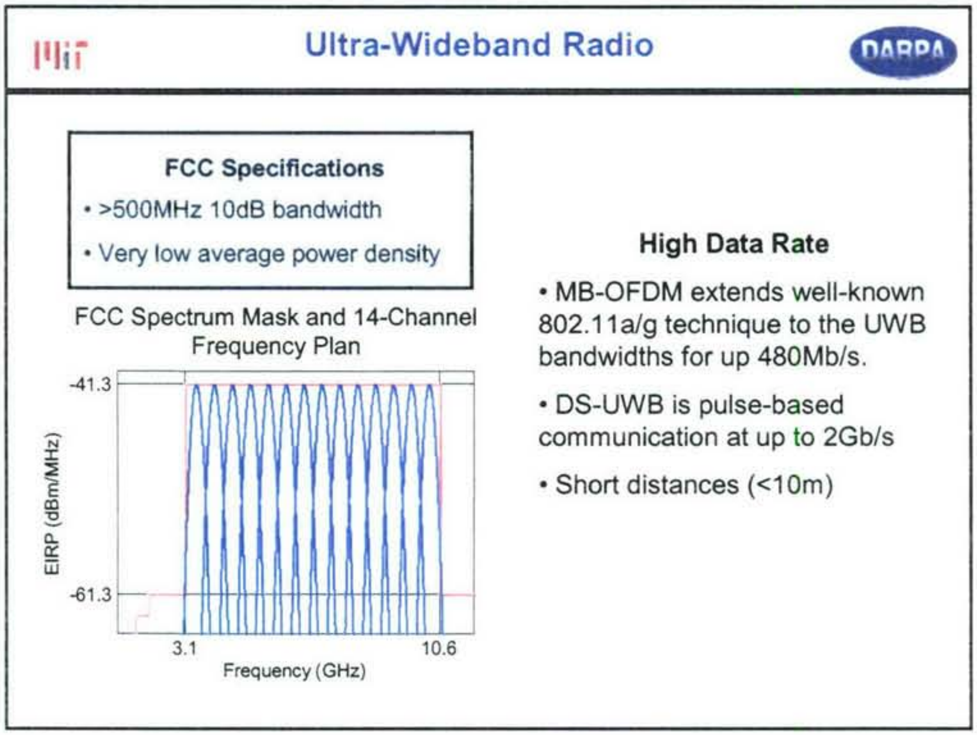


Table of Contents

- System Motivation
- Time-Interleaved ADC
- Digital baseband processor
- Mixed-signal optimum energy point
- Conclusion



Ultra-Wideband Radio

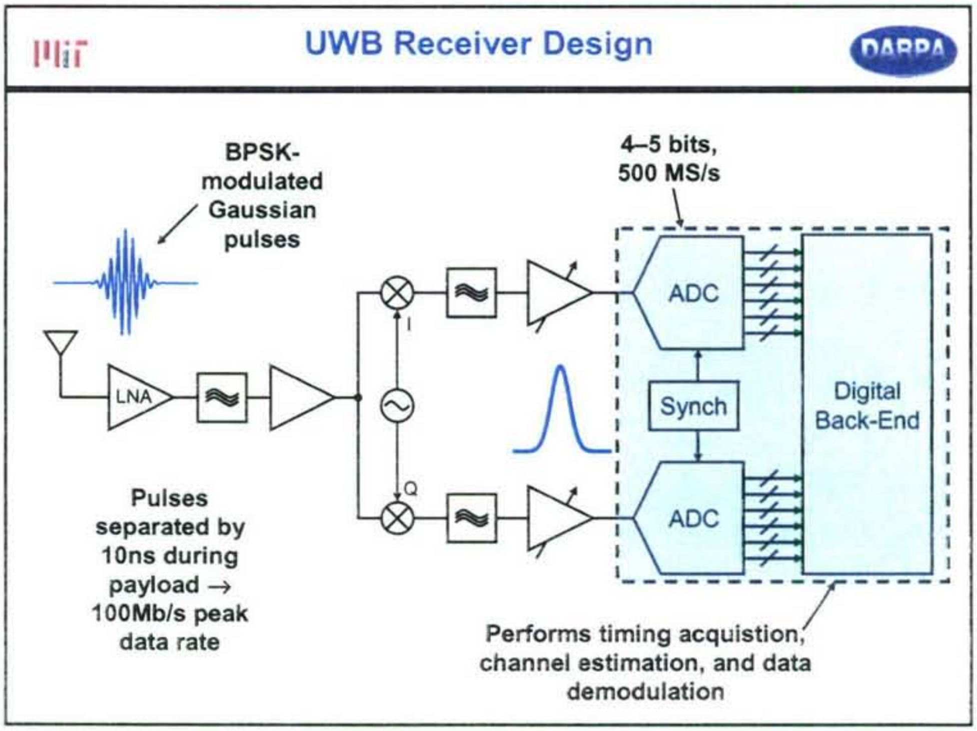
FCC Specifications

- >500MHz 10dB bandwidth
- Very low average power density

FCC Spectrum Mask and 14-Channel Frequency Plan

High Data Rate

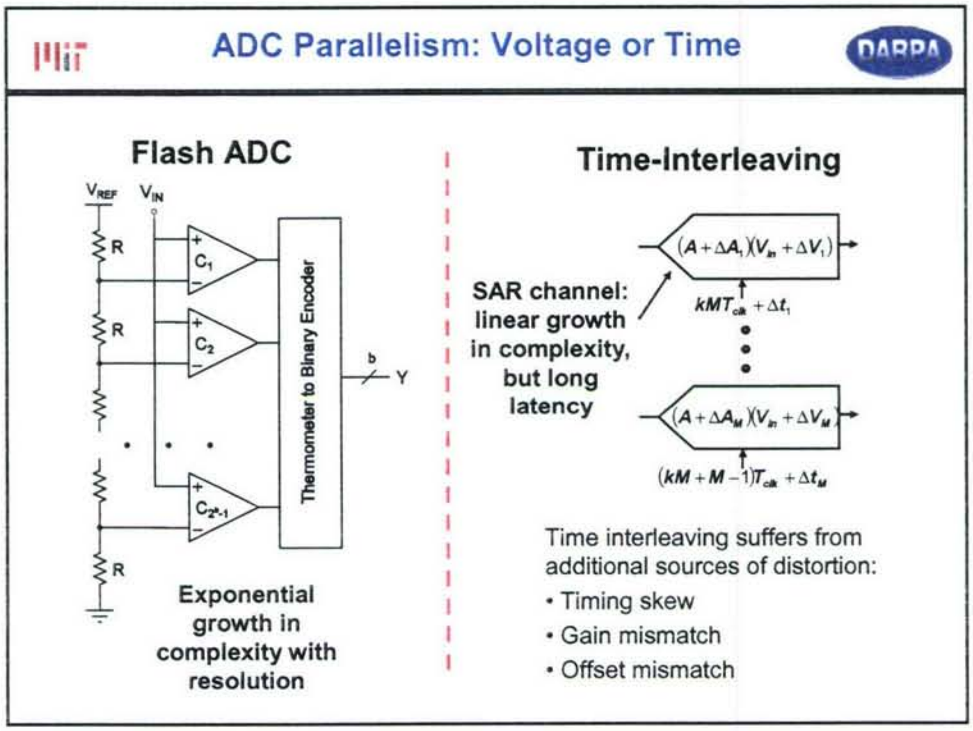
- MB-OFDM extends well-known 802.11a/g technique to the UWB bandwidths for up 480Mb/s.
- DS-UWB is pulse-based communication at up to 2Gb/s
- Short distances (<10m)



UWB Receiver Design

Pulses separated by 10ns during payload → 100Mb/s peak data rate

Performs timing acquistion, channel estimation, and data demodulation



ADC Parallelism: Voltage or Time

Flash ADC

V_{REF}
V_{IN}
R
R
R
C_1
C_2
C_{2^b-1}
Thermometer to Binary Encoder
b
Y

Exponential growth in complexity with resolution

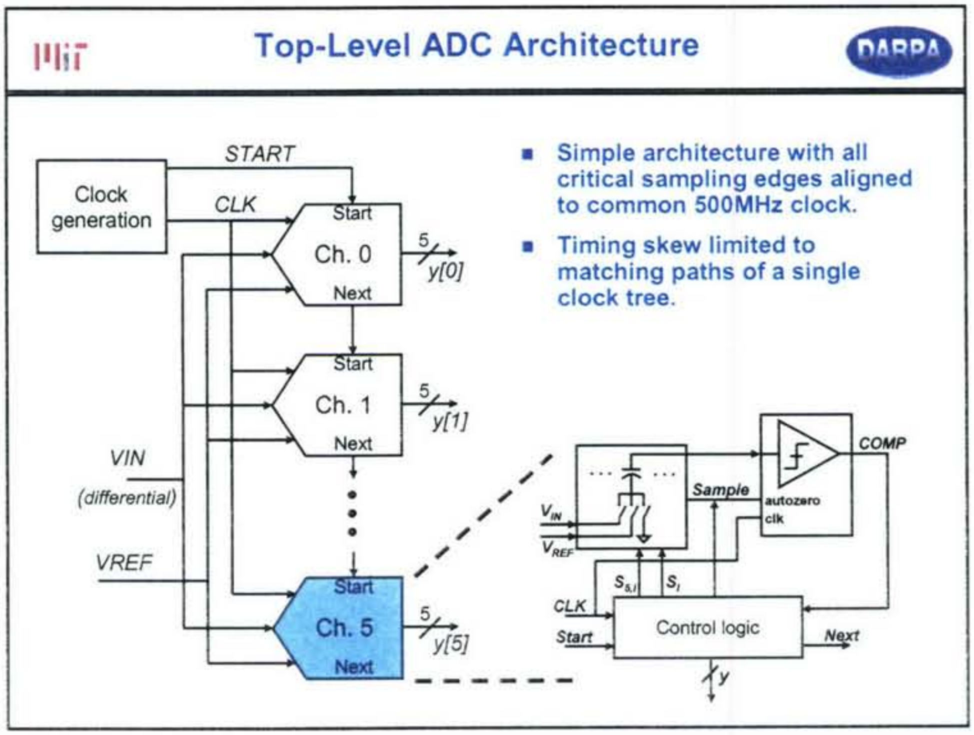
Time-Interleaving

SAR channel: linear growth in complexity, but long latency

$(A+\Delta A_1)(V_{in}+\Delta V_1)$
$kMT_{clk}+\Delta t_1$
$(A+\Delta A_M)(V_{in}+\Delta V_M)$
$(kM+M-1)T_{clk}+\Delta t_M$

Time interleaving suffers from additional sources of distortion:

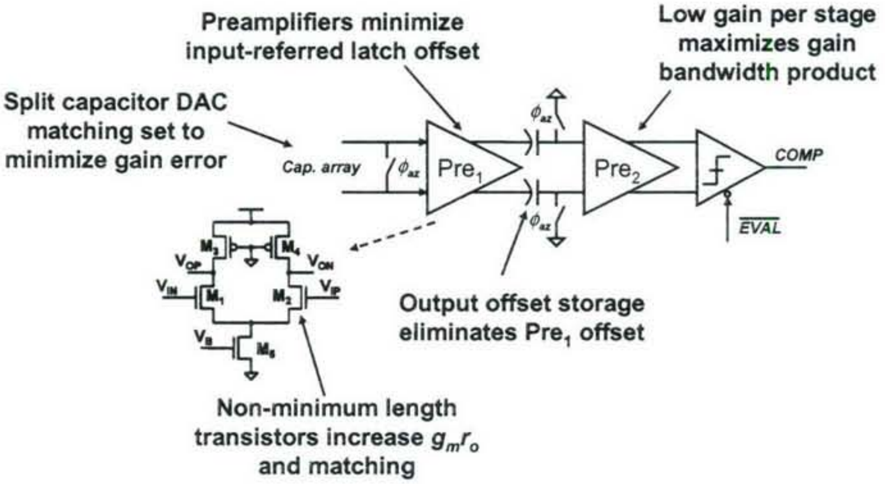
- Timing skew
- Gain mismatch
- Offset mismatch



Top-Level ADC Architecture

- Simple architecture with all critical sampling edges aligned to common 500MHz clock.
- Timing skew limited to matching paths of a single clock tree.

Circuit Details: Comparator


Preamplifiers minimize input-referred latch offset

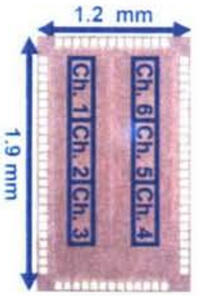
Low gain per stage maximizes gain bandwidth product

Split capacitor DAC matching set to minimize gain error

Output offset storage eliminates Pre$_1$ offset

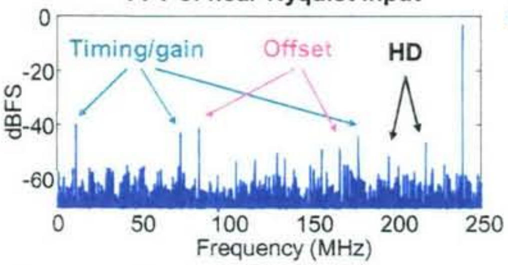
Non-minimum length transistors increase $g_m r_o$ and matching

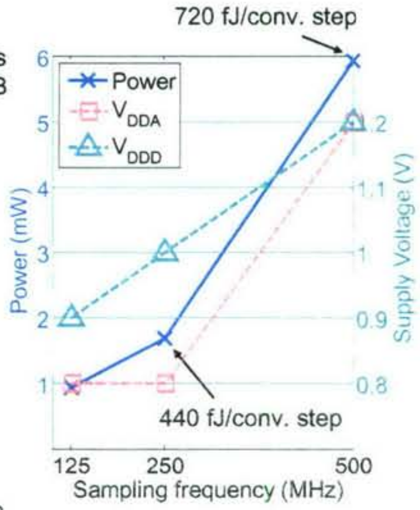
ADC Measurements

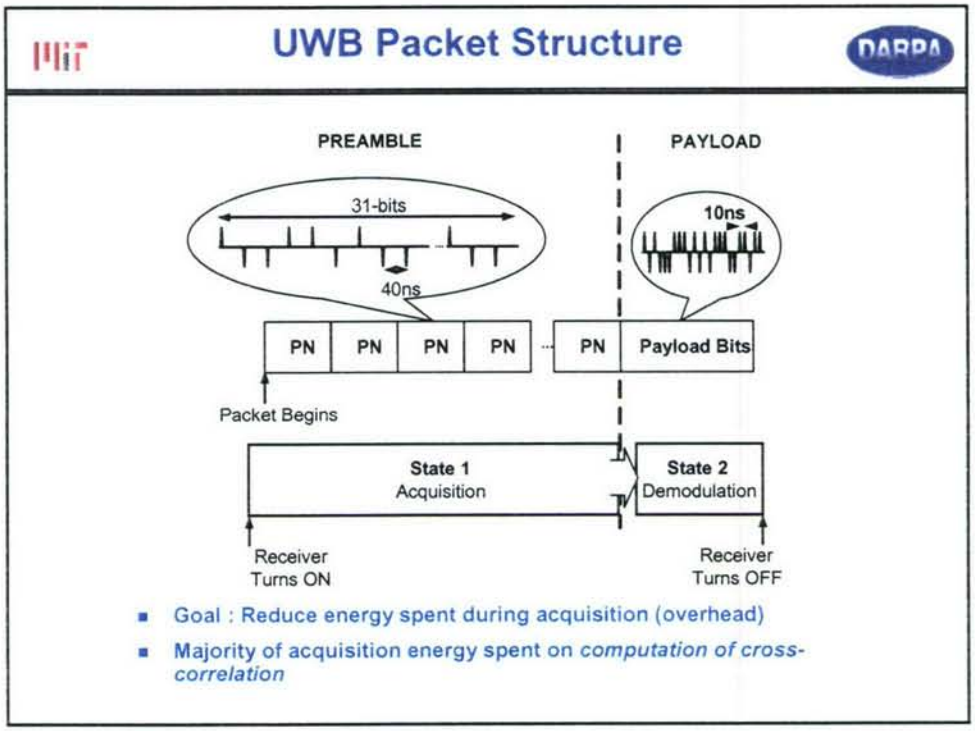


- TI 65nm CMOS process
- INL<0.2, DNL<0.25 LSB
- ENOB = 4.5 (DC), 4 (Nyquist)
- Complete results in [Ginsburg, VLSI 2006]

FFT of near-Nyquist input

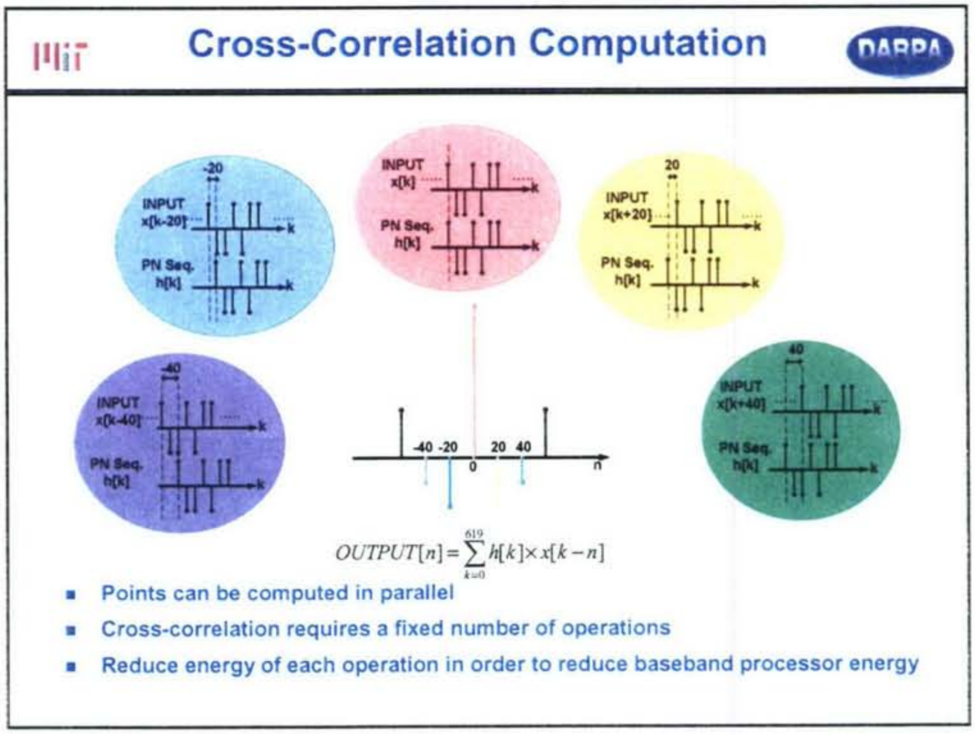






UWB Packet Structure

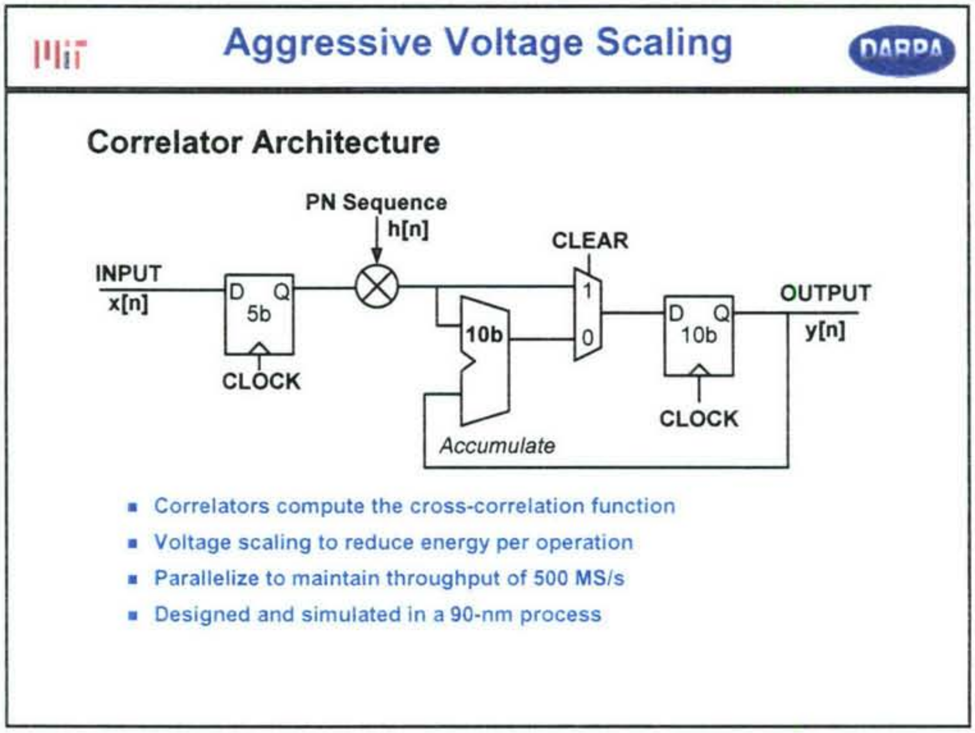
- Goal : Reduce energy spent during acquisition (overhead)
- Majority of acquisition energy spent on *computation of cross-correlation*



Cross-Correlation Computation

$$OUTPUT[n] = \sum_{k=0}^{619} h[k] \times x[k-n]$$

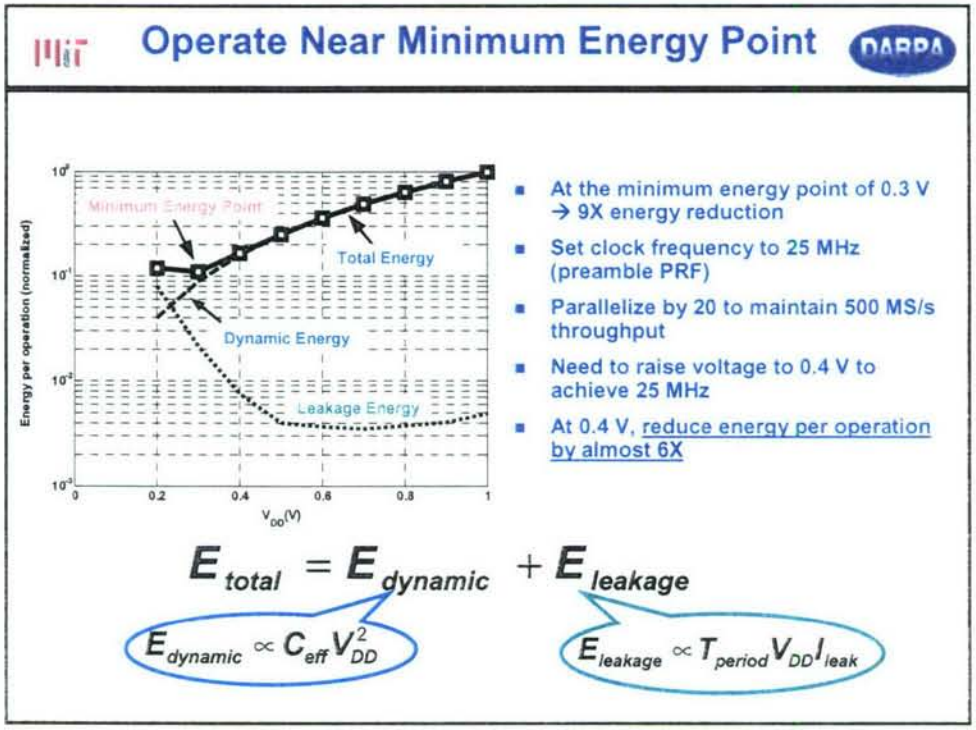
- Points can be computed in parallel
- Cross-correlation requires a fixed number of operations
- Reduce energy of each operation in order to reduce baseband processor energy



Aggressive Voltage Scaling

Correlator Architecture

- Correlators compute the cross-correlation function
- Voltage scaling to reduce energy per operation
- Parallelize to maintain throughput of 500 MS/s
- Designed and simulated in a 90-nm process



Operate Near Minimum Energy Point

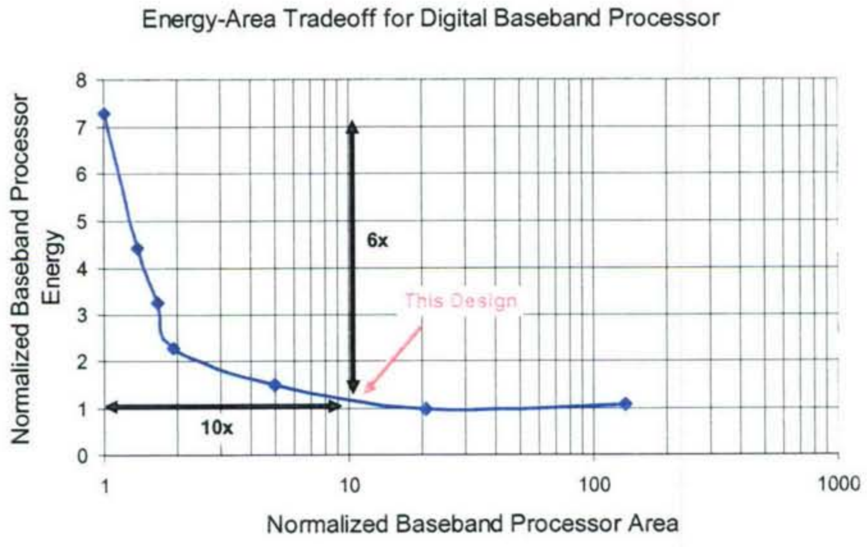
- At the minimum energy point of 0.3 V → 9X energy reduction
- Set clock frequency to 25 MHz (preamble PRF)
- Parallelize by 20 to maintain 500 MS/s throughput
- Need to raise voltage to 0.4 V to achieve 25 MHz
- At 0.4 V, reduce energy per operation by almost 6X

$$E_{total} = E_{dynamic} + E_{leakage}$$

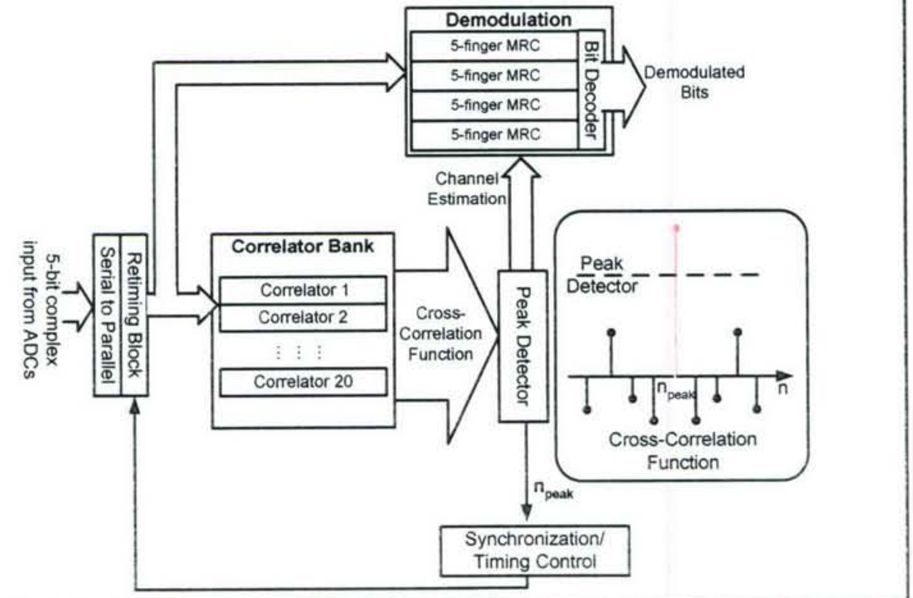
$$E_{dynamic} \propto C_{eff} V_{DD}^2$$

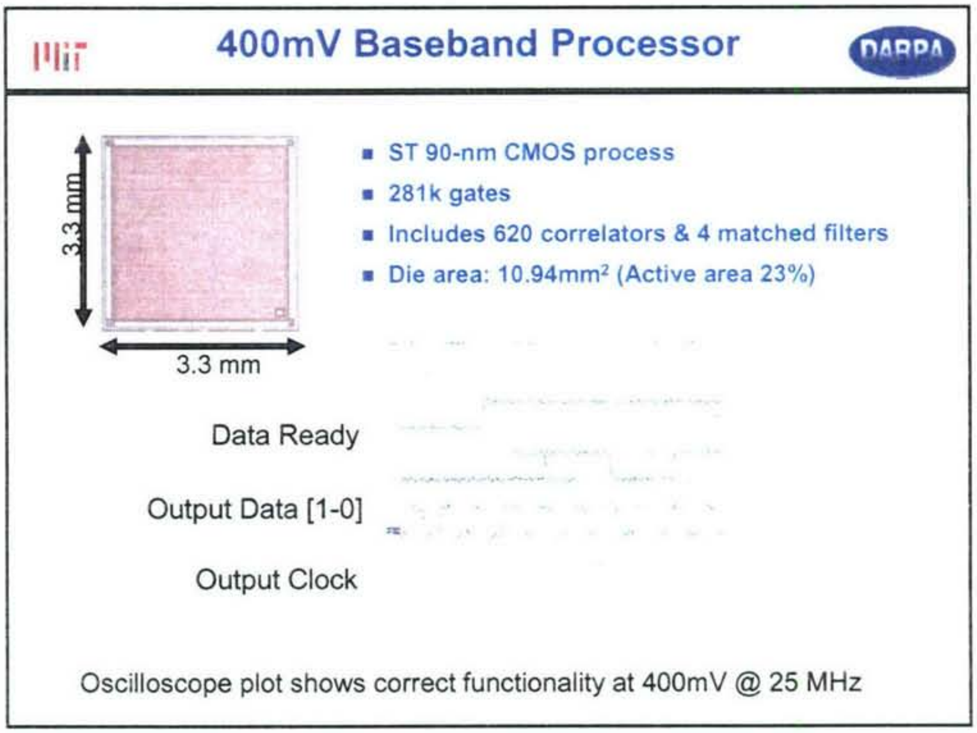
$$E_{leakage} \propto T_{period} V_{DD} I_{leak}$$

Energy-Area Tradeoff



Baseband Architecture



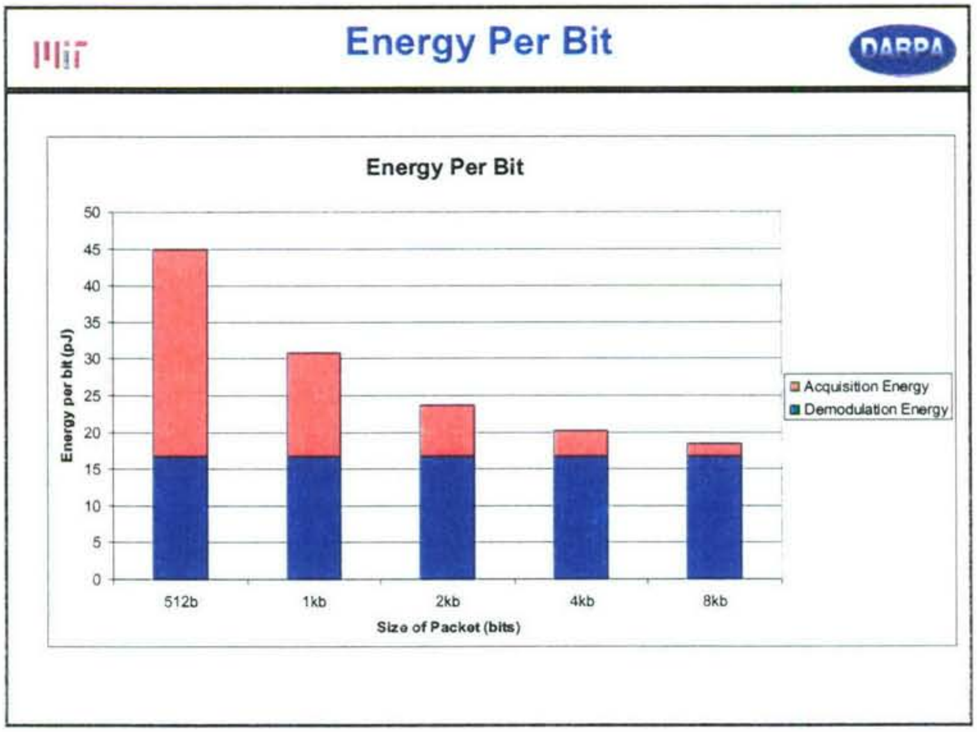


400mV Baseband Processor

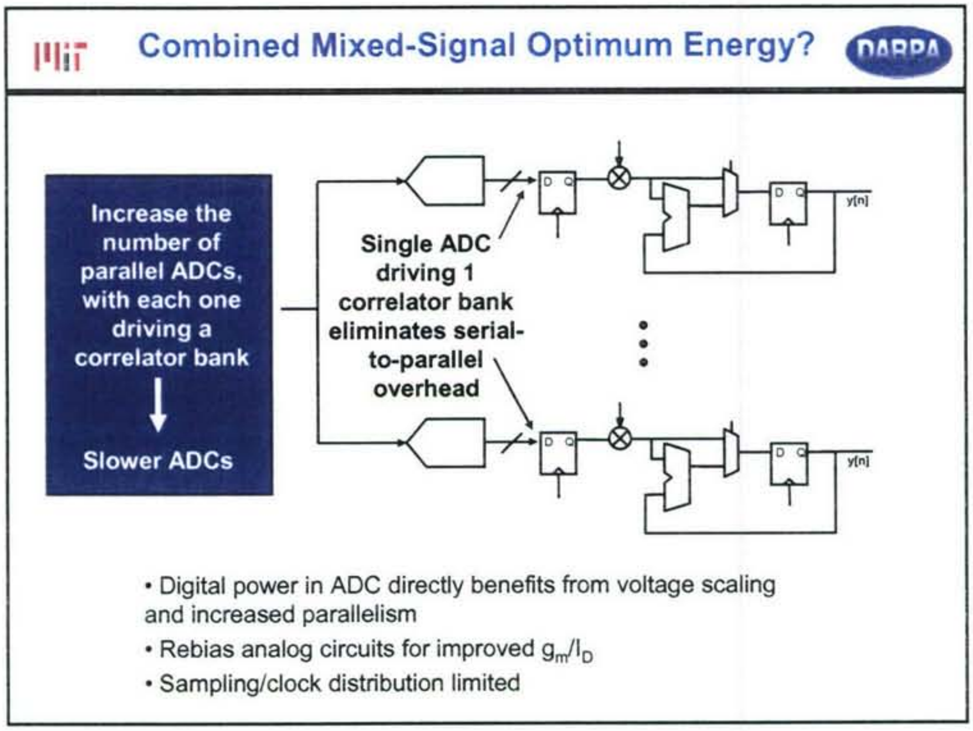
- ST 90-nm CMOS process
- 281k gates
- Includes 620 correlators & 4 matched filters
- Die area: 10.94mm² (Active area 23%)

3.3 mm
3.3 mm
Data Ready
Output Data [1-0]
Output Clock

Oscilloscope plot shows correct functionality at 400mV @ 25 MHz



Energy Per Bit



Combined Mixed-Signal Optimum Energy?

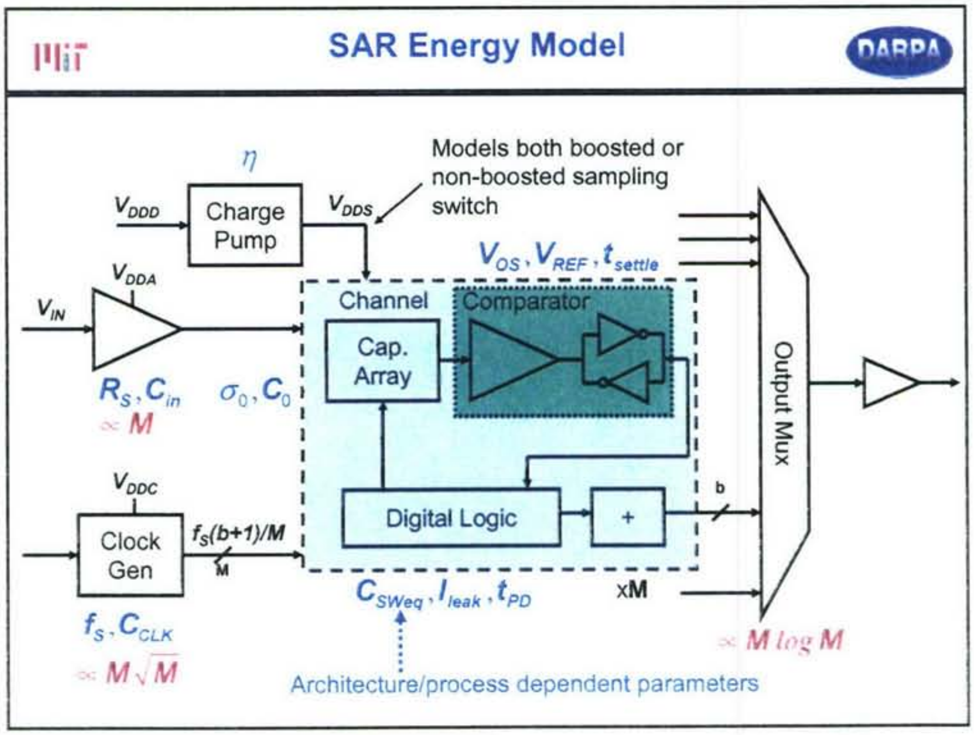
Increase the number of parallel ADCs, with each one driving a correlator bank

↓

Slower ADCs

Single ADC driving 1 correlator bank eliminates serial-to-parallel overhead

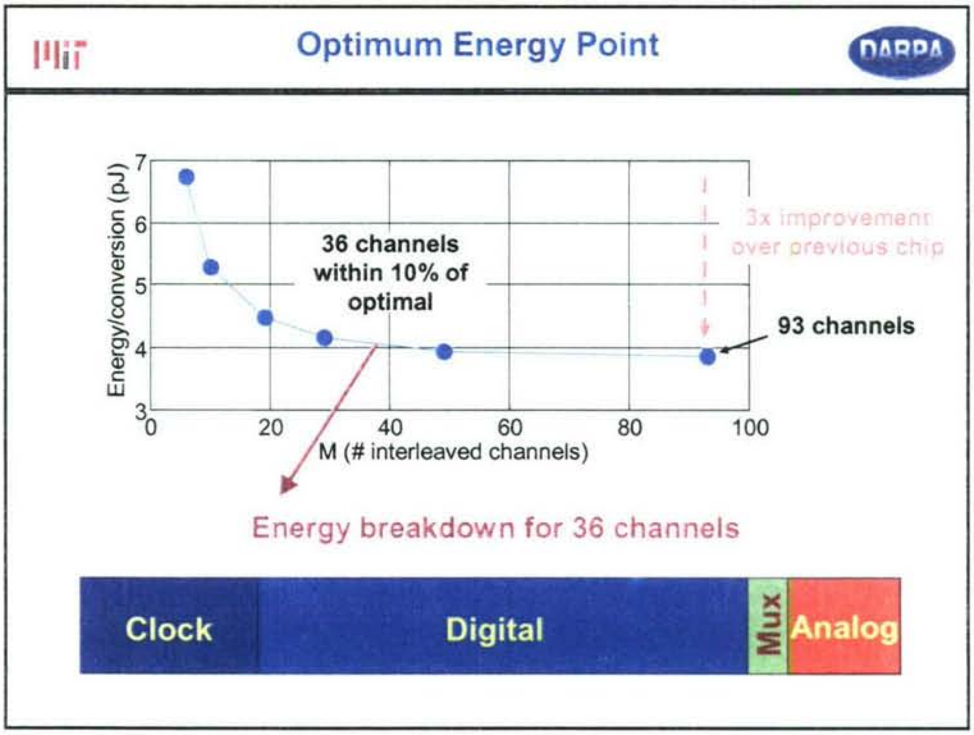
- Digital power in ADC directly benefits from voltage scaling and increased parallelism
- Rebias analog circuits for improved g_m/I_D
- Sampling/clock distribution limited



SAR Energy Model

Models both boosted or non-boosted sampling switch

Architecture/process dependent parameters



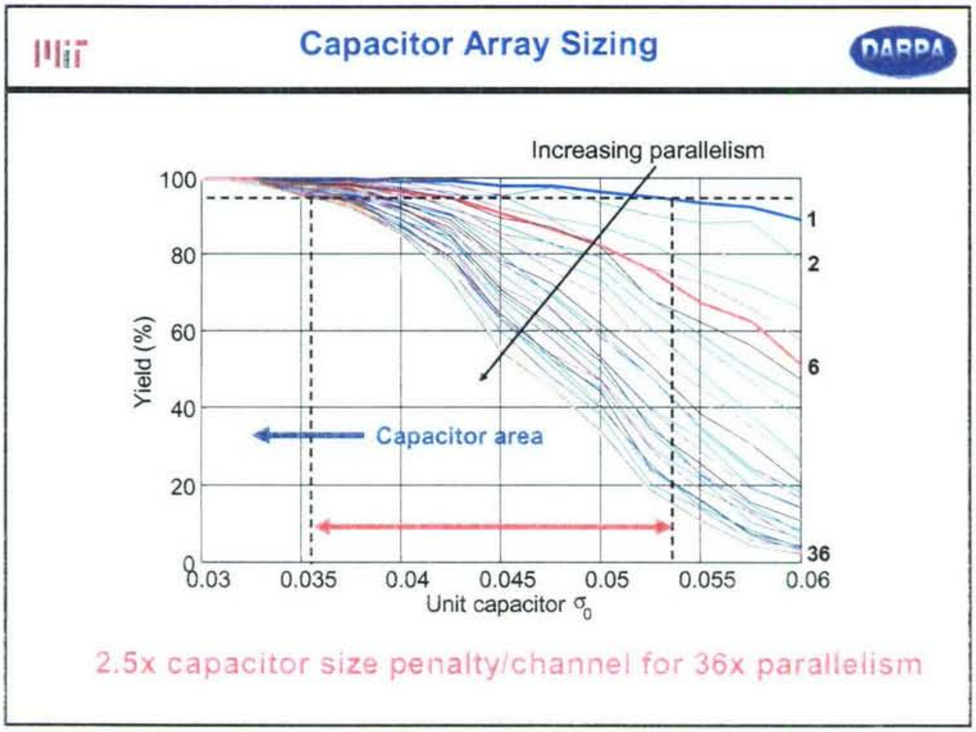
Optimum Energy Point

36 channels within 10% of optimal

3x improvement over previous chip

93 channels

Energy breakdown for 36 channels

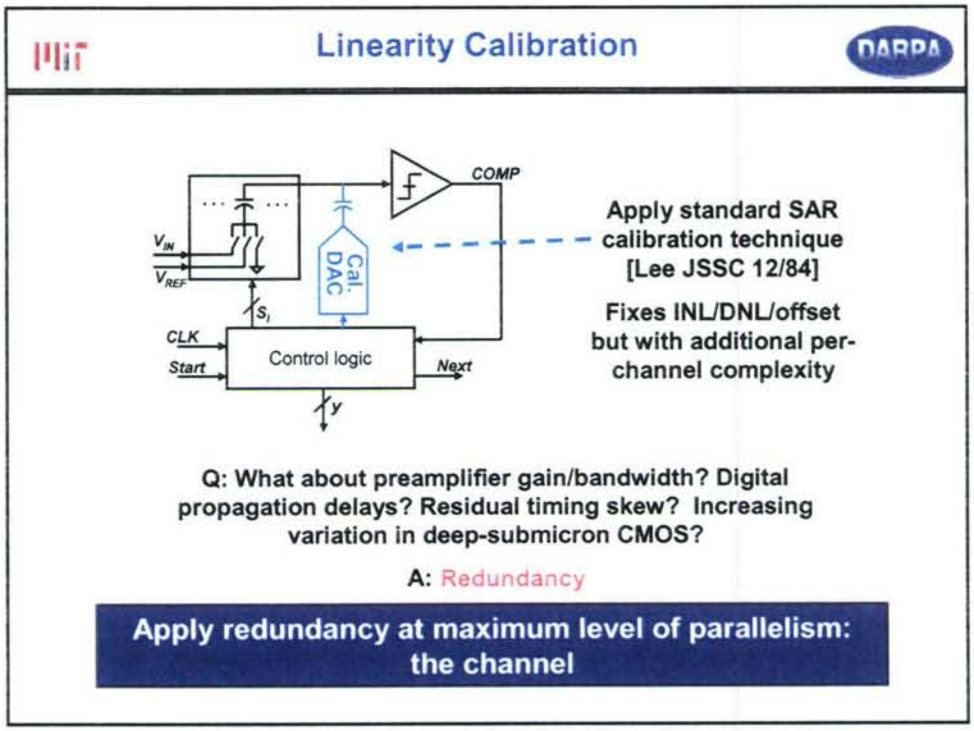


Capacitor Array Sizing

Increasing parallelism

Capacitor area

2.5x capacitor size penalty/channel for 36x parallelism



Linearity Calibration

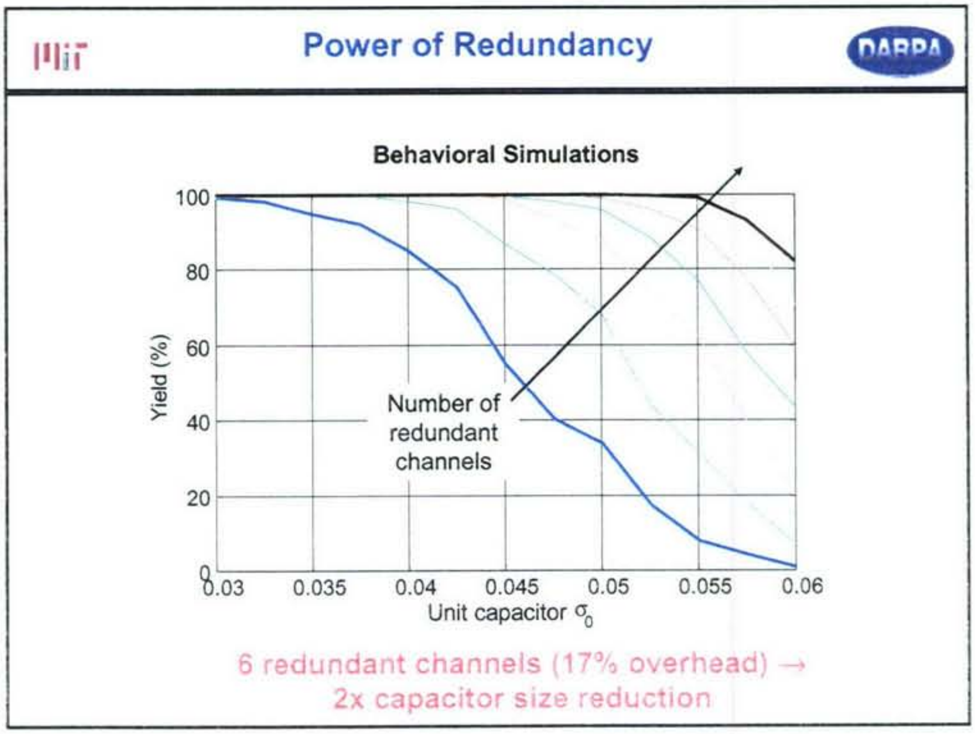
Apply standard SAR calibration technique [Lee JSSC 12/84]

Fixes INL/DNL/offset but with additional per-channel complexity

Q: What about preamplifier gain/bandwidth? Digital propagation delays? Residual timing skew? Increasing variation in deep-submicron CMOS?

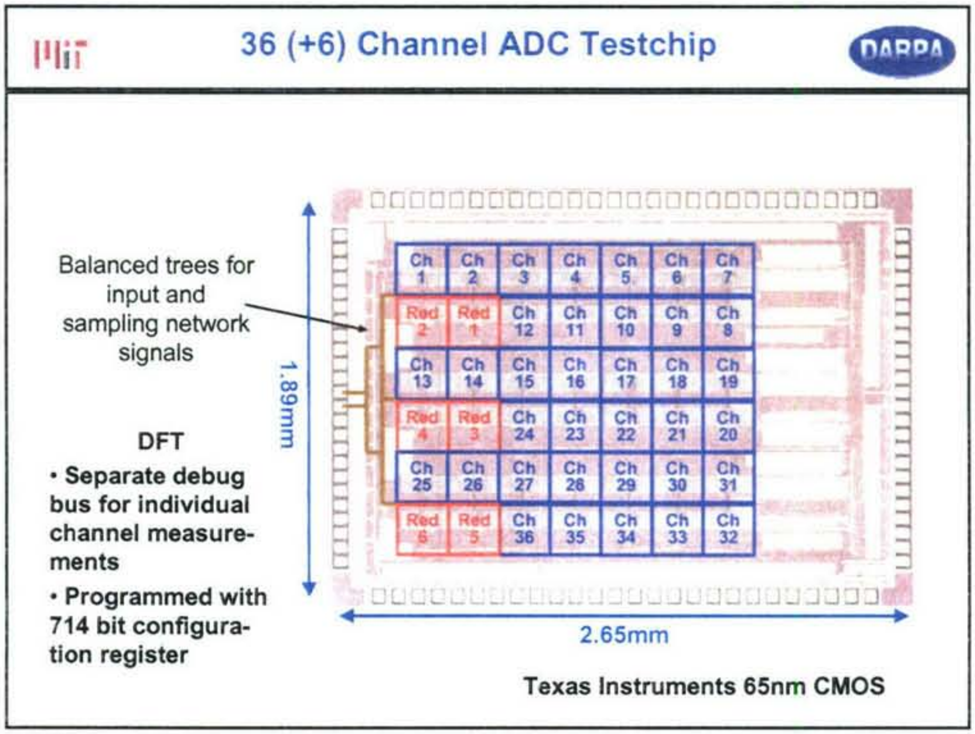
A: Redundancy

Apply redundancy at maximum level of parallelism: the channel



Power of Redundancy

6 redundant channels (17% overhead) → 2x capacitor size reduction



36 (+6) Channel ADC Testchip

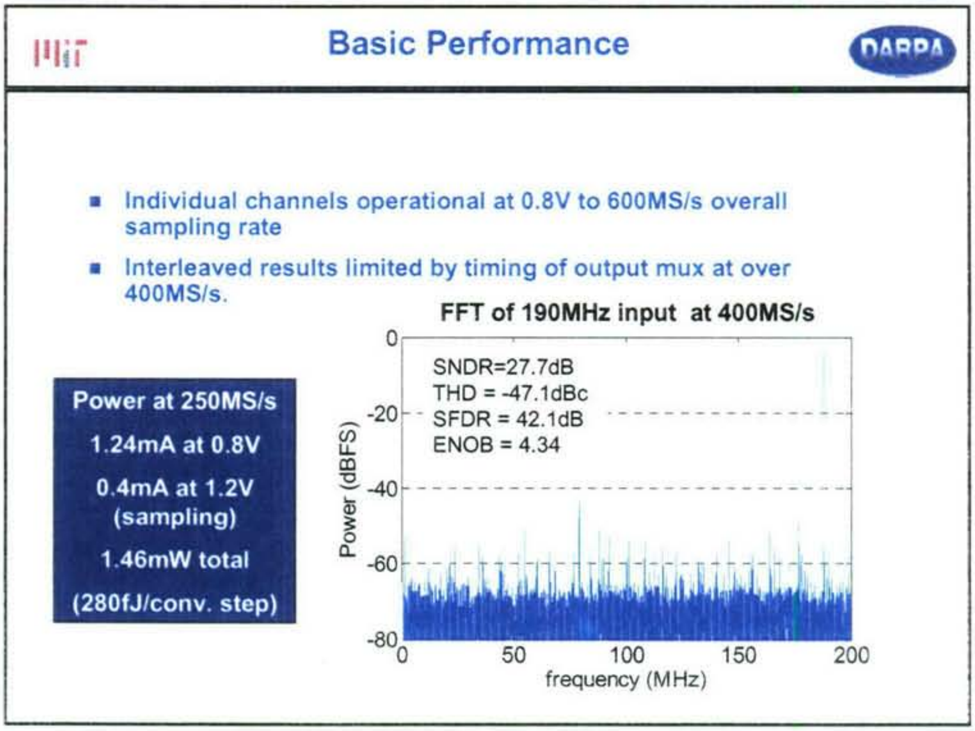
Balanced trees for input and sampling network signals

DFT

- **Separate debug bus for individual channel measurements**
- **Programmed with 714 bit configuration register**

Ch 1	Ch 2	Ch 3	Ch 4	Ch 5	Ch 6	Ch 7
Red 2	Red 1	Ch 12	Ch 11	Ch 10	Ch 9	Ch 8
Ch 13	Ch 14	Ch 15	Ch 16	Ch 17	Ch 18	Ch 19
Red 4	Red 3	Ch 24	Ch 23	Ch 22	Ch 21	Ch 20
Ch 25	Ch 26	Ch 27	Ch 28	Ch 29	Ch 30	Ch 31
Red 6	Red 5	Ch 36	Ch 35	Ch 34	Ch 33	Ch 32

Texas Instruments 65nm CMOS



Basic Performance

- Individual channels operational at 0.8V to 600MS/s overall sampling rate
- Interleaved results limited by timing of output mux at over 400MS/s.

Power at 250MS/s

1.24mA at 0.8V

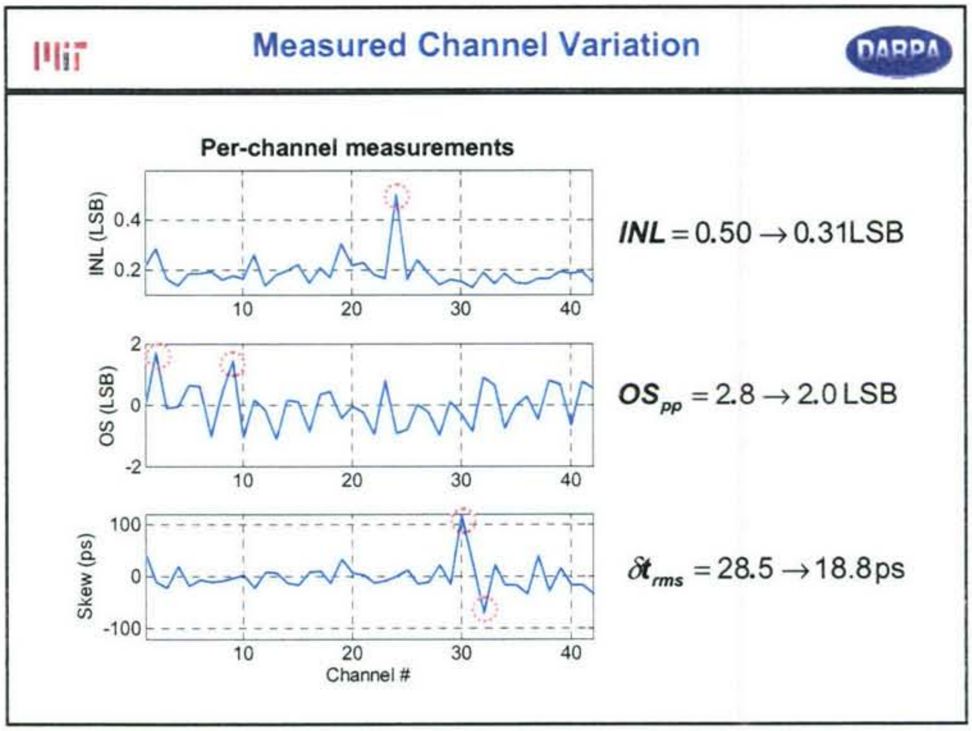
0.4mA at 1.2V (sampling)

1.46mW total

(280fJ/conv. step)

FFT of 190MHz input at 400MS/s

SNDR=27.7dB
THD = -47.1dBc
SFDR = 42.1dB
ENOB = 4.34



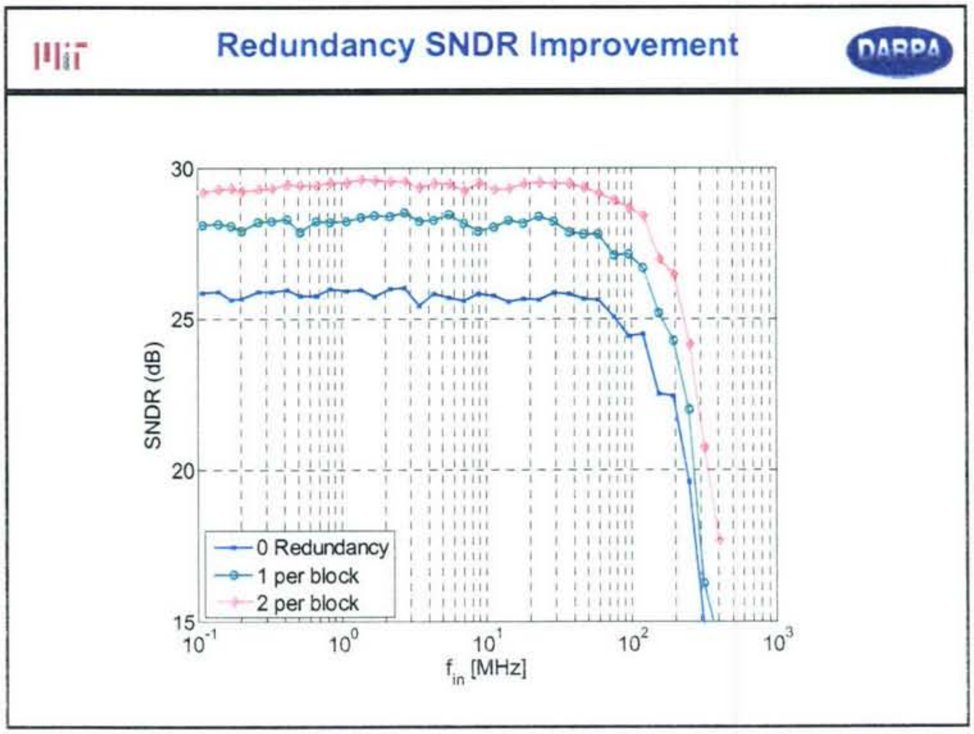
Measured Channel Variation

Per-channel measurements

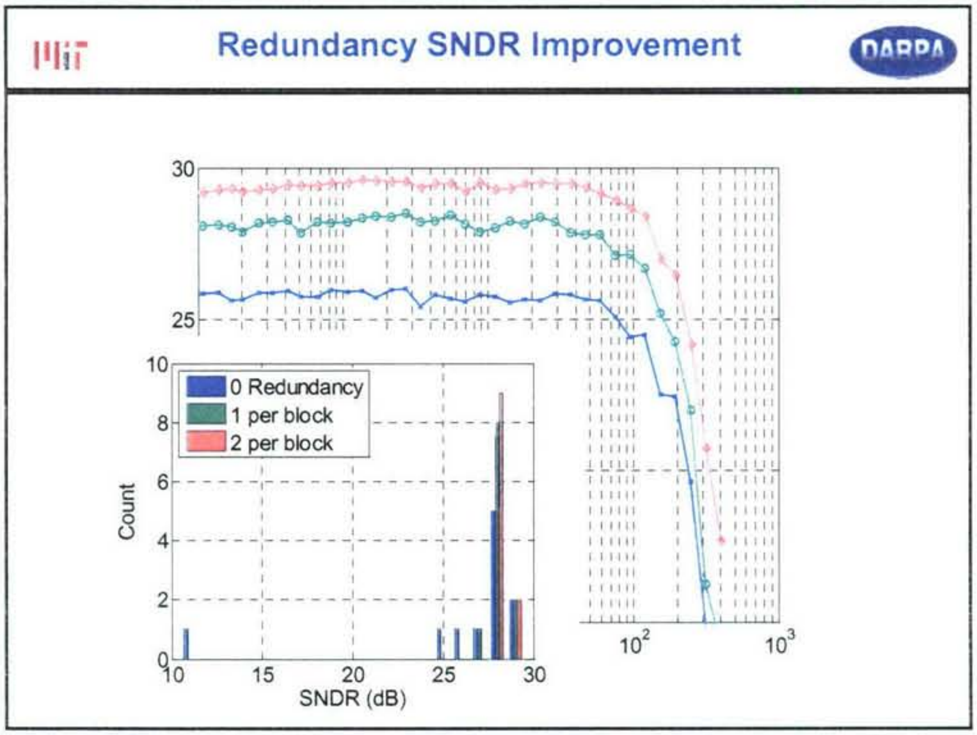
$INL = 0.50 \rightarrow 0.31\text{LSB}$

$OS_{pp} = 2.8 \rightarrow 2.0\,\text{LSB}$

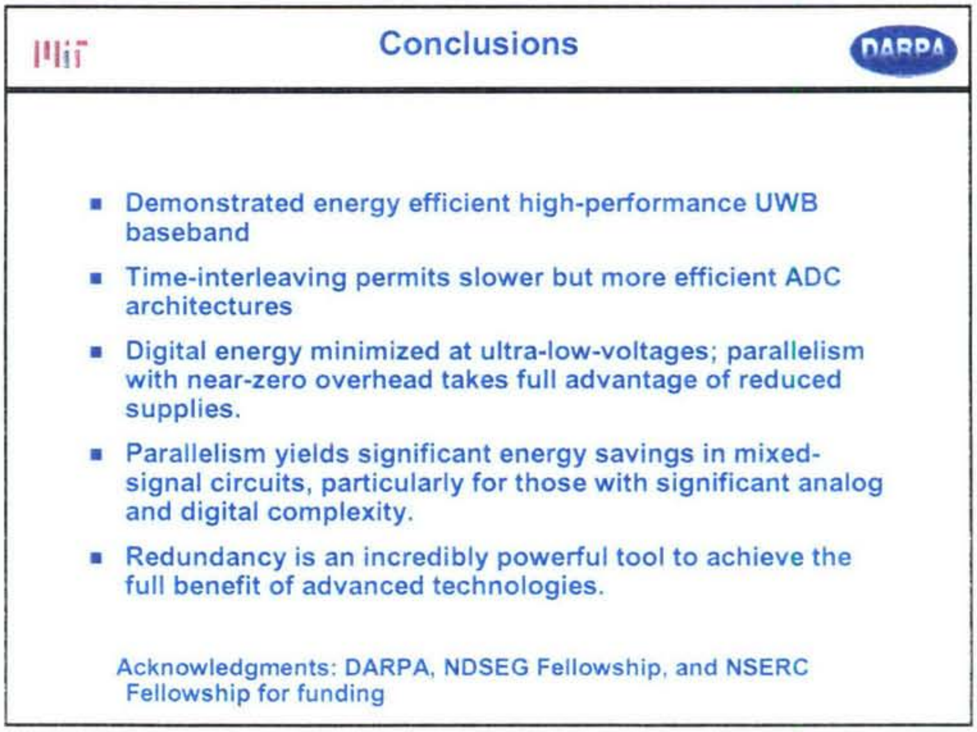
$\delta t_{rms} = 28.5 \rightarrow 18.8\,\text{ps}$



Redundancy SNDR Improvement



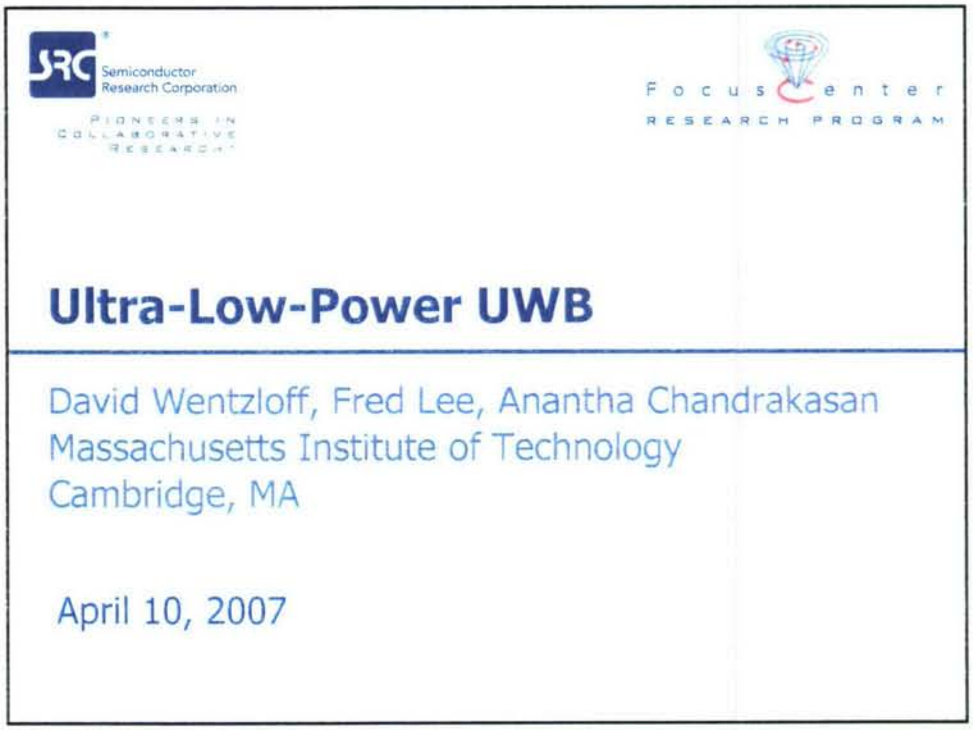
Redundancy SNDR Improvement



Conclusions

- Demonstrated energy efficient high-performance UWB baseband
- Time-interleaving permits slower but more efficient ADC architectures
- Digital energy minimized at ultra-low-voltages; parallelism with near-zero overhead takes full advantage of reduced supplies.
- Parallelism yields significant energy savings in mixed-signal circuits, particularly for those with significant analog and digital complexity.
- Redundancy is an incredibly powerful tool to achieve the full benefit of advanced technologies.

Acknowledgments: DARPA, NDSEG Fellowship, and NSERC Fellowship for funding



SRC® Semiconductor Research Corporation

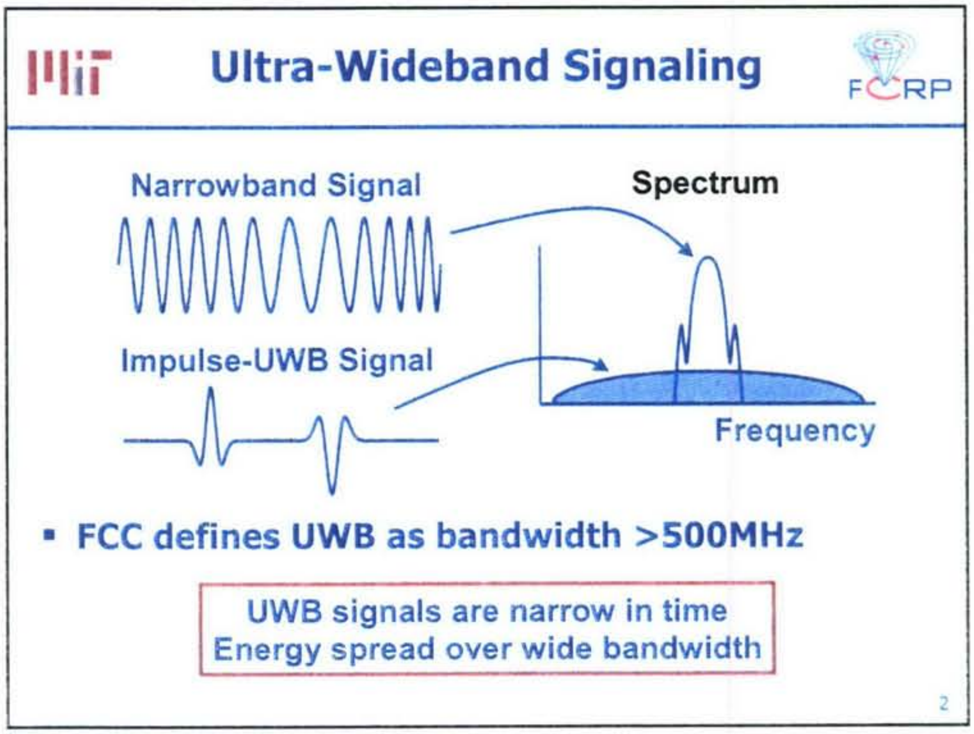
Pioneers in Collaborative Research®

Focus Center Research Program

Ultra-Low-Power UWB

David Wentzloff, Fred Lee, Anantha Chandrakasan
Massachusetts Institute of Technology
Cambridge, MA

April 10, 2007



MIT

Ultra-Wideband Signaling

FCRP

Narrowband Signal

Spectrum

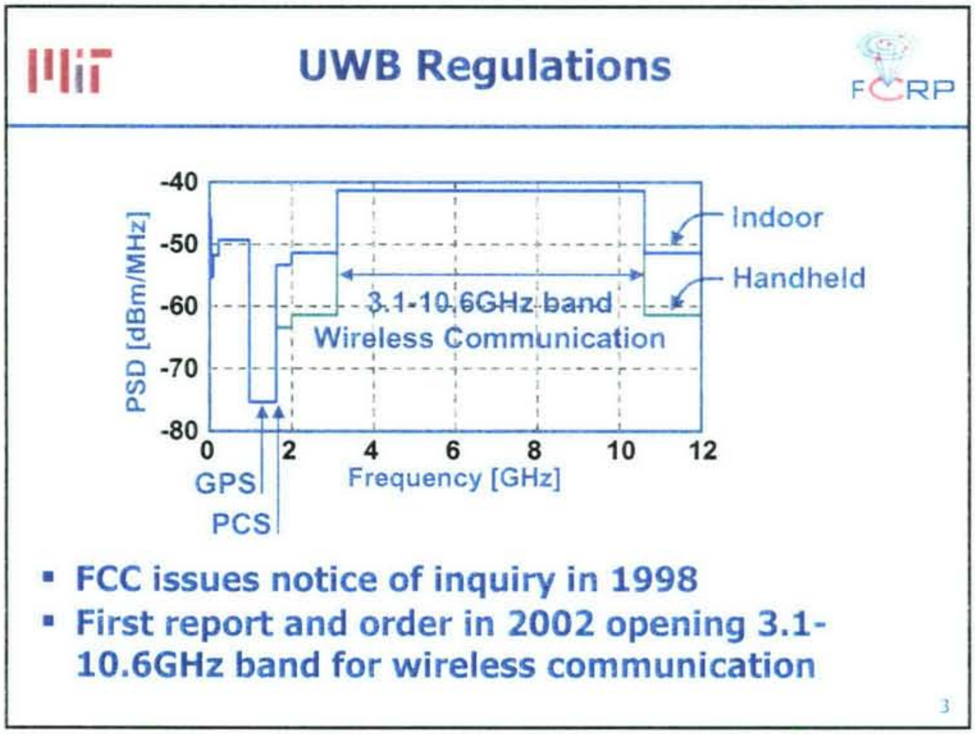
Impulse-UWB Signal

Frequency

- FCC defines UWB as bandwidth >500MHz

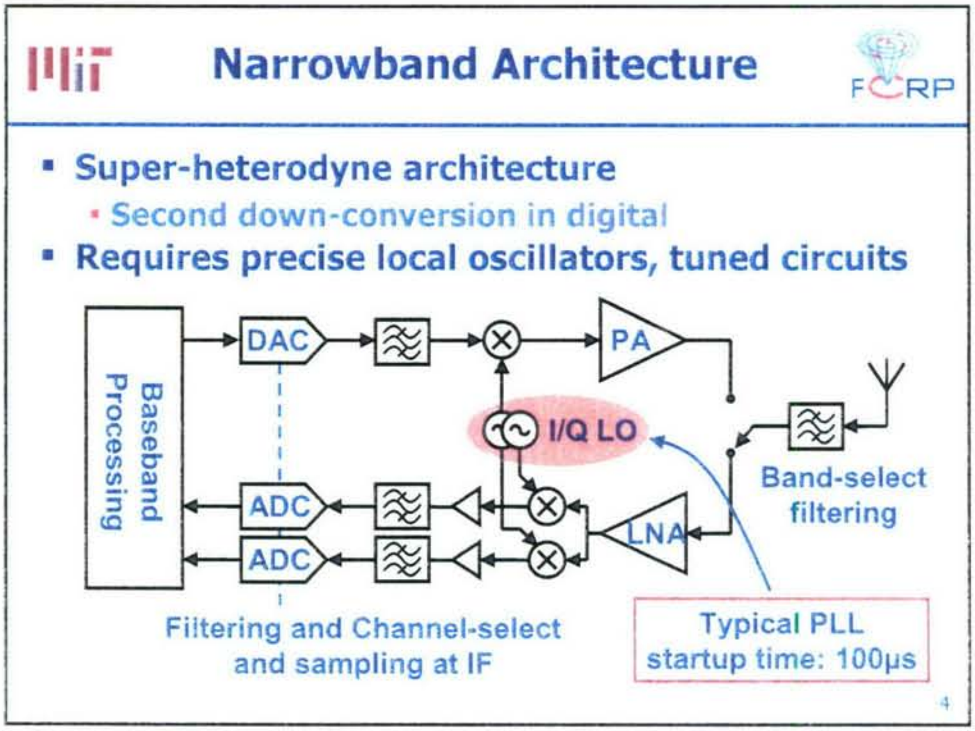
UWB signals are narrow in time
Energy spread over wide bandwidth

2



UWB Regulations

- FCC issues notice of inquiry in 1998
- First report and order in 2002 opening 3.1-10.6GHz band for wireless communication

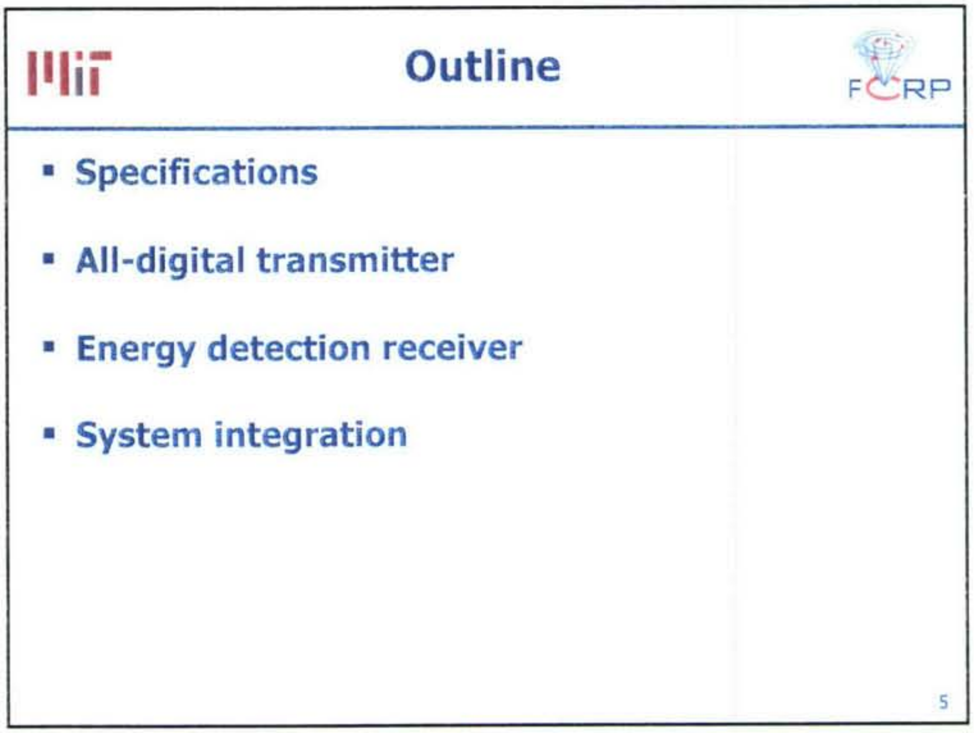


Narrowband Architecture

- Super-heterodyne architecture
  - Second down-conversion in digital
- Requires precise local oscillators, tuned circuits

Filtering and Channel-select and sampling at IF

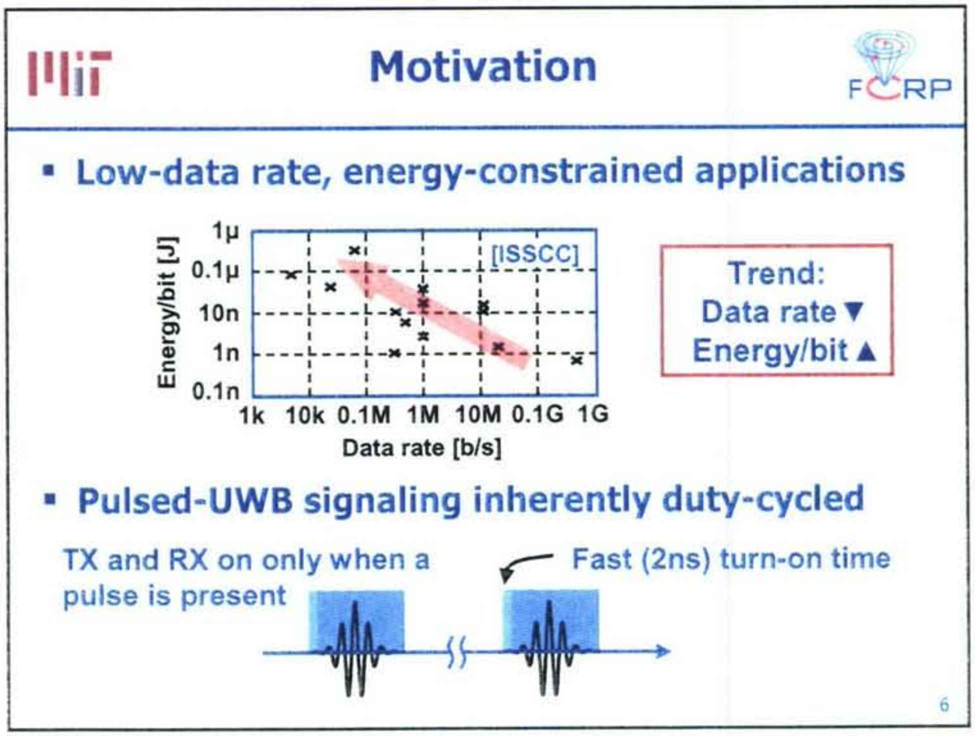
Typical PLL startup time: 100μs



Outline

- Specifications
- All-digital transmitter
- Energy detection receiver
- System integration

5



Motivation

- Low-data rate, energy-constrained applications

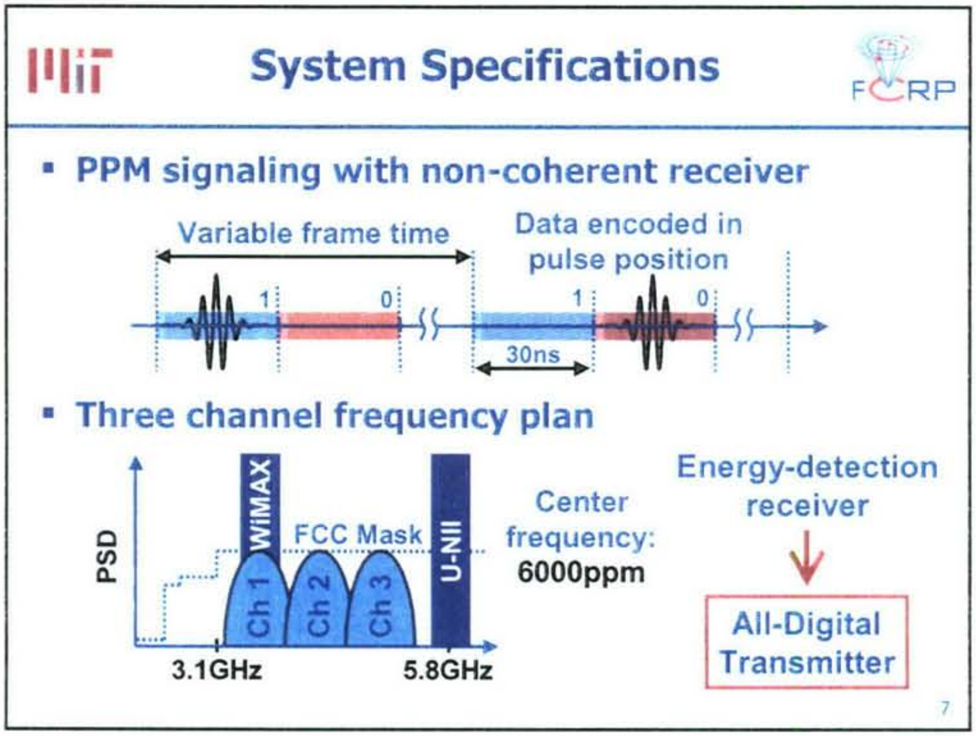
Trend:
Data rate ▼
Energy/bit ▲

- Pulsed-UWB signaling inherently duty-cycled

TX and RX on only when a pulse is present

Fast (2ns) turn-on time

6



System Specifications

- PPM signaling with non-coherent receiver

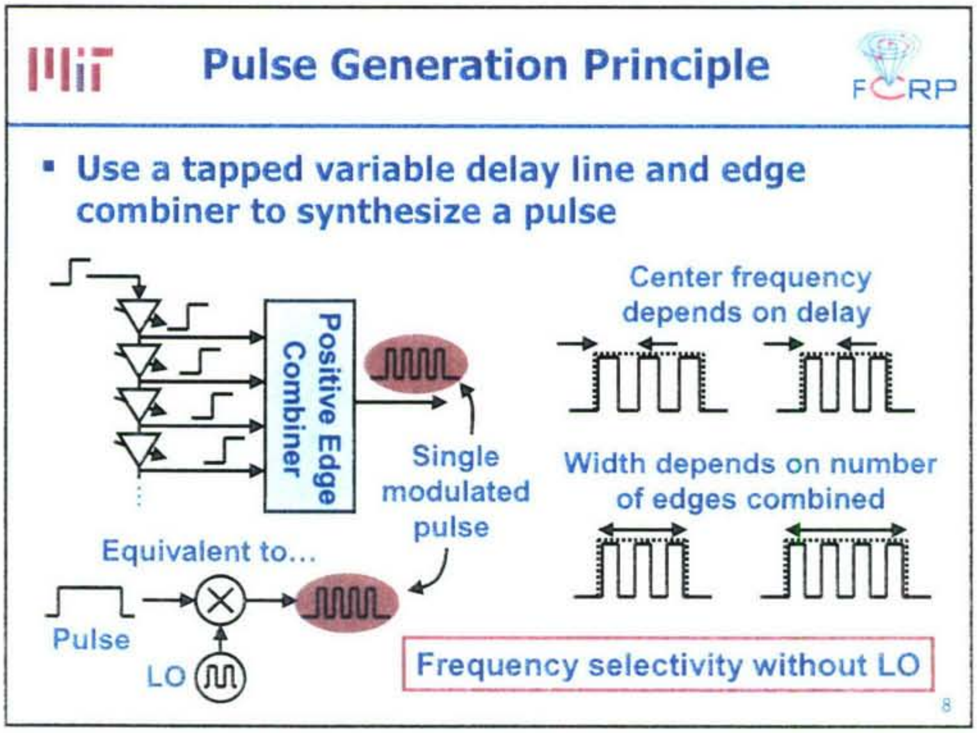
- Three channel frequency plan

Center frequency: 6000ppm

Energy-detection receiver

All-Digital Transmitter

7



Pulse Generation Principle

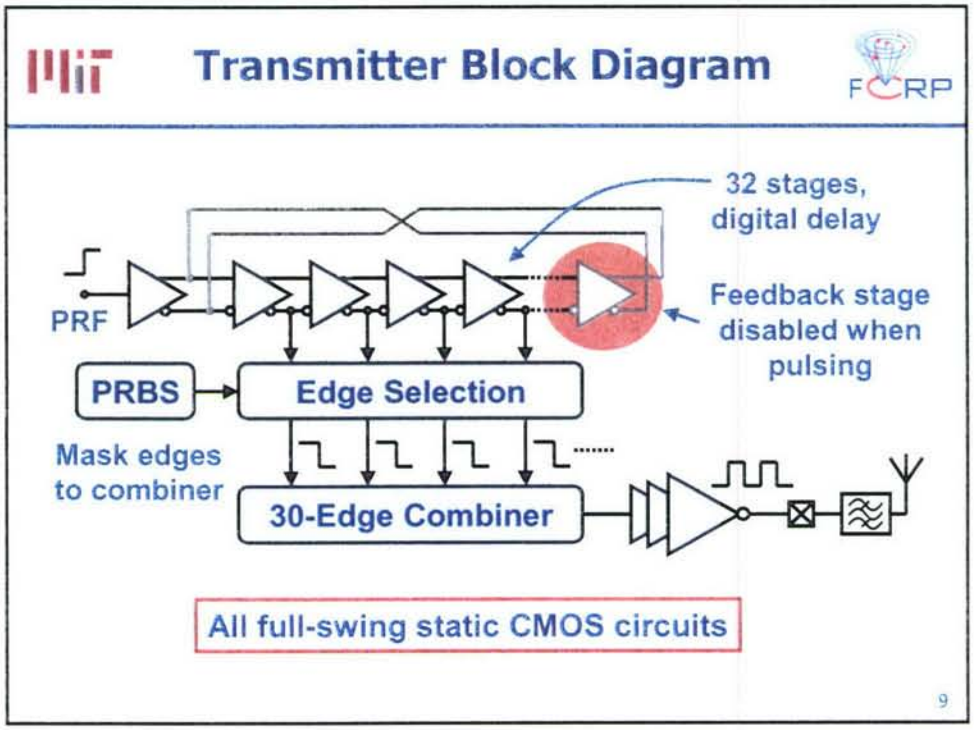
- Use a tapped variable delay line and edge combiner to synthesize a pulse

Center frequency depends on delay

Width depends on number of edges combined

Frequency selectivity without LO

8



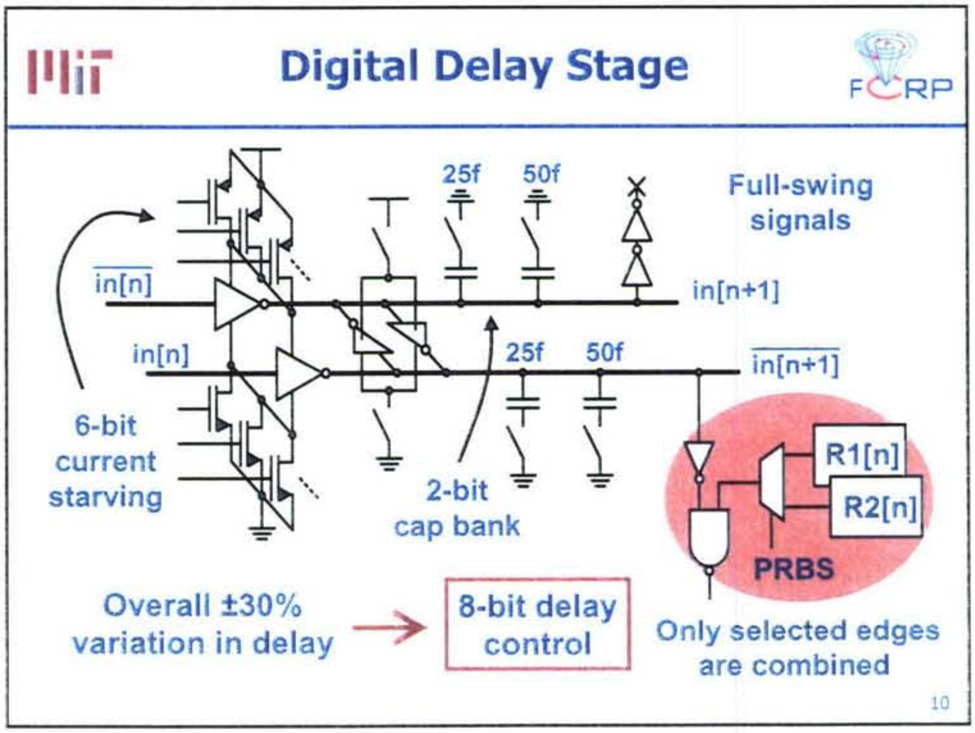
Transmitter Block Diagram

32 stages, digital delay

Feedback stage disabled when pulsing

Mask edges to combiner

All full-swing static CMOS circuits



Digital Delay Stage

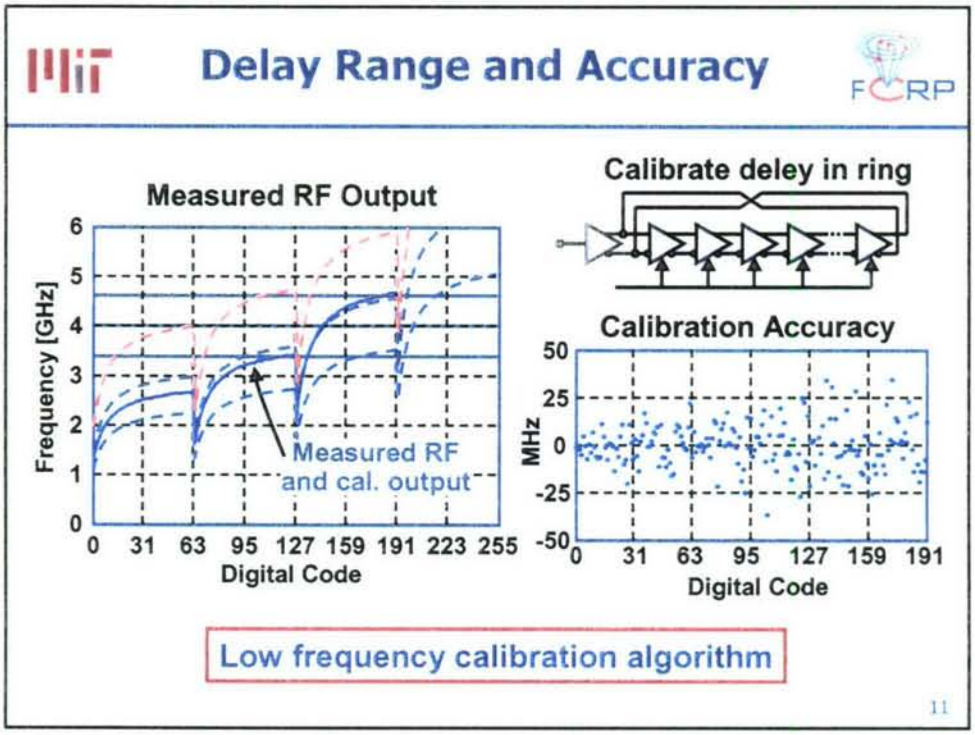
Full-swing signals

6-bit current starving

2-bit cap bank

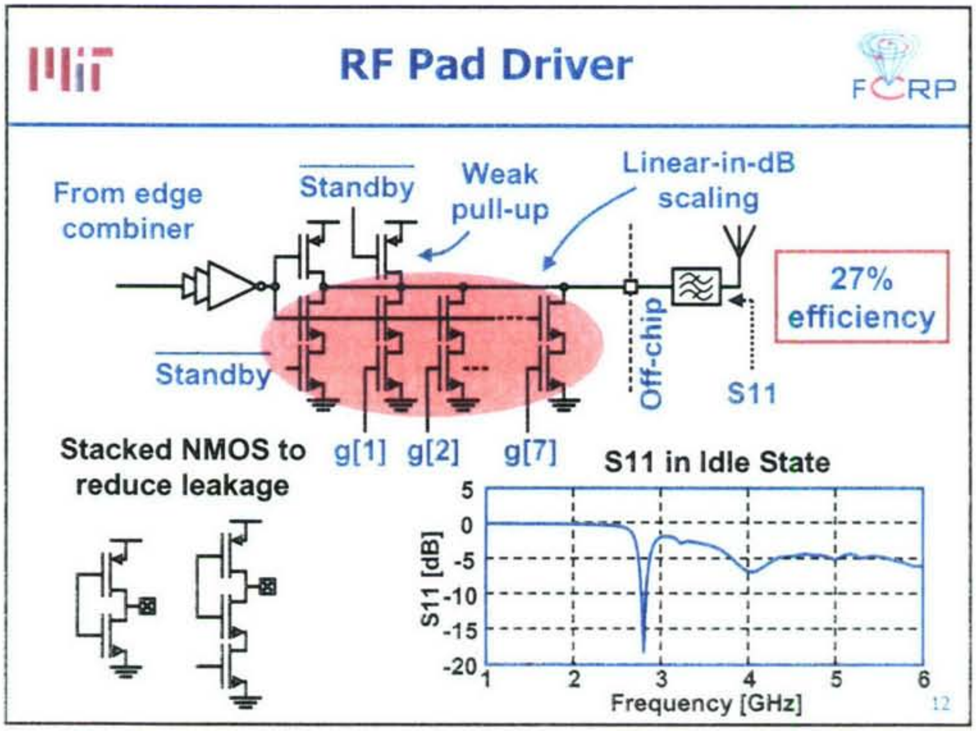
Only selected edges are combined

Overall ±30% variation in delay → 8-bit delay control



Delay Range and Accuracy

Low frequency calibration algorithm

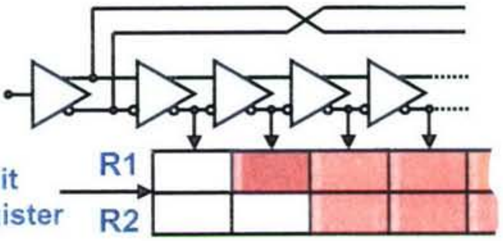


RF Pad Driver

27% efficiency

Stacked NMOS to reduce leakage

Delay-Based BPSK Scrambling

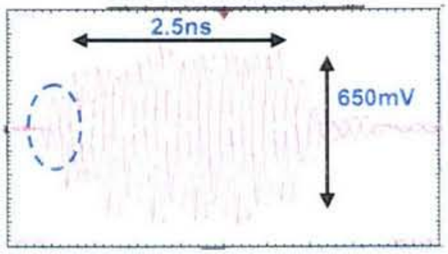


PRBS bit selects register

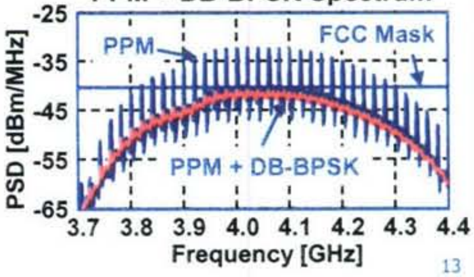
Per-stage delay is ½ RF period

Mask values offset by 1 bit

DB-BPSK Pulses

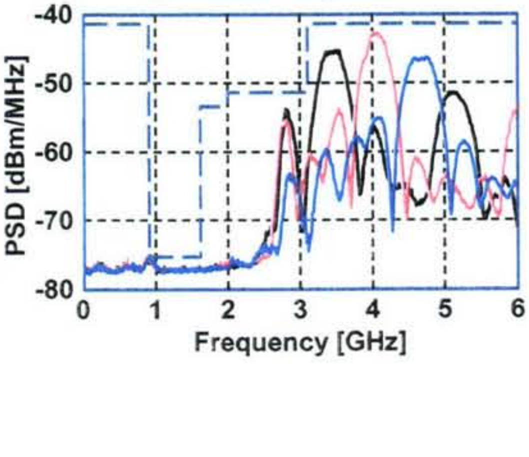


PPM + DB-BPSK Spectrum

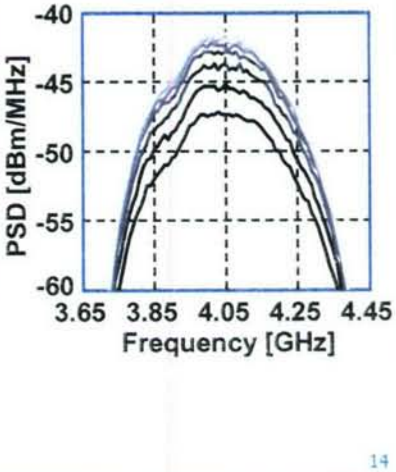


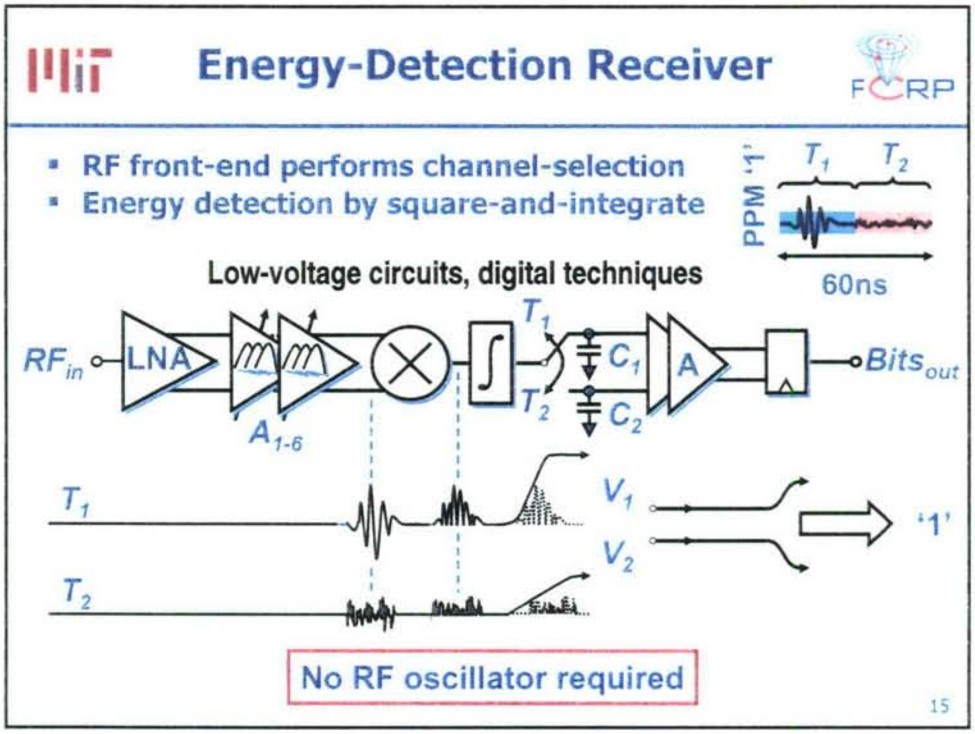
Measured Spectrum

3-Channel Spectrum



CH2 Gain Settings



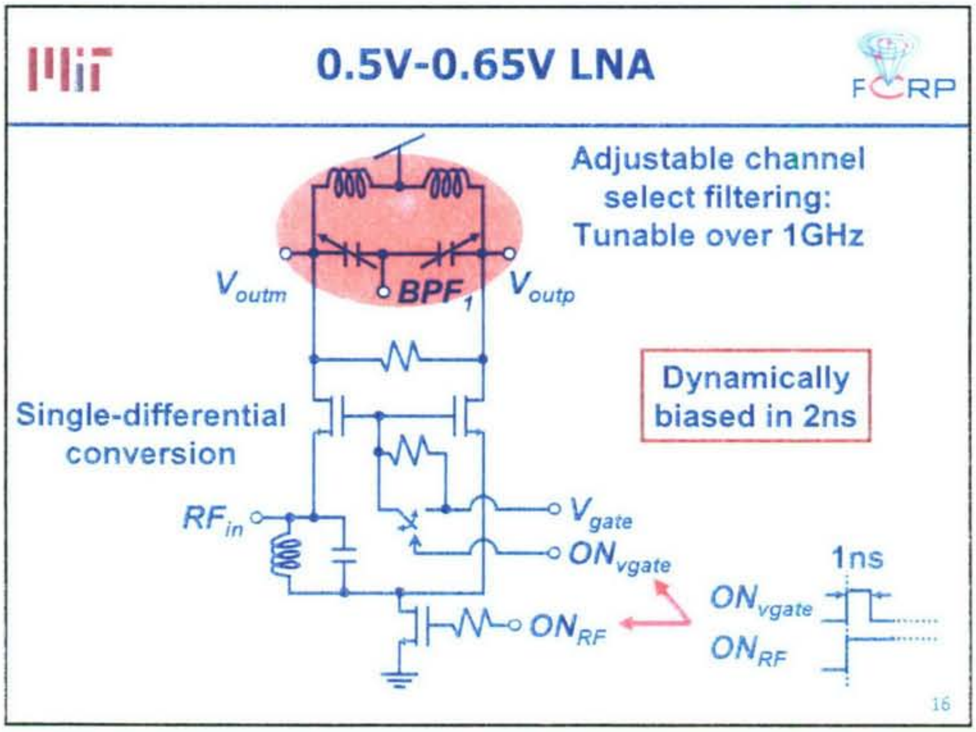


Energy-Detection Receiver

- RF front-end performs channel-selection
- Energy detection by square-and-integrate

Low-voltage circuits, digital techniques

No RF oscillator required

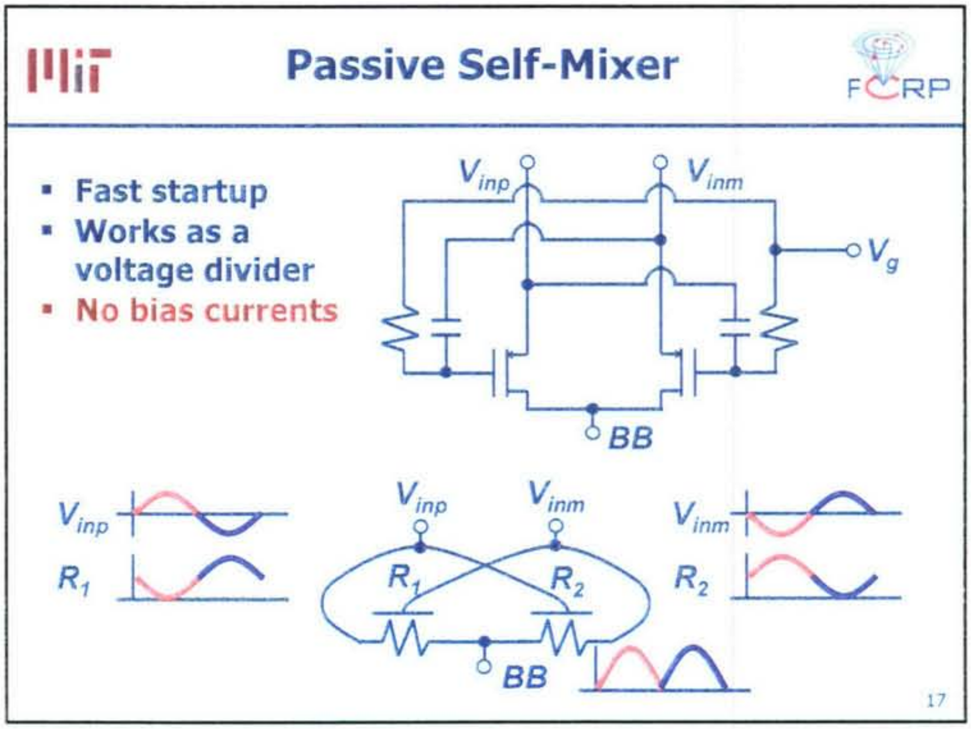


0.5V-0.65V LNA

Adjustable channel select filtering: Tunable over 1GHz

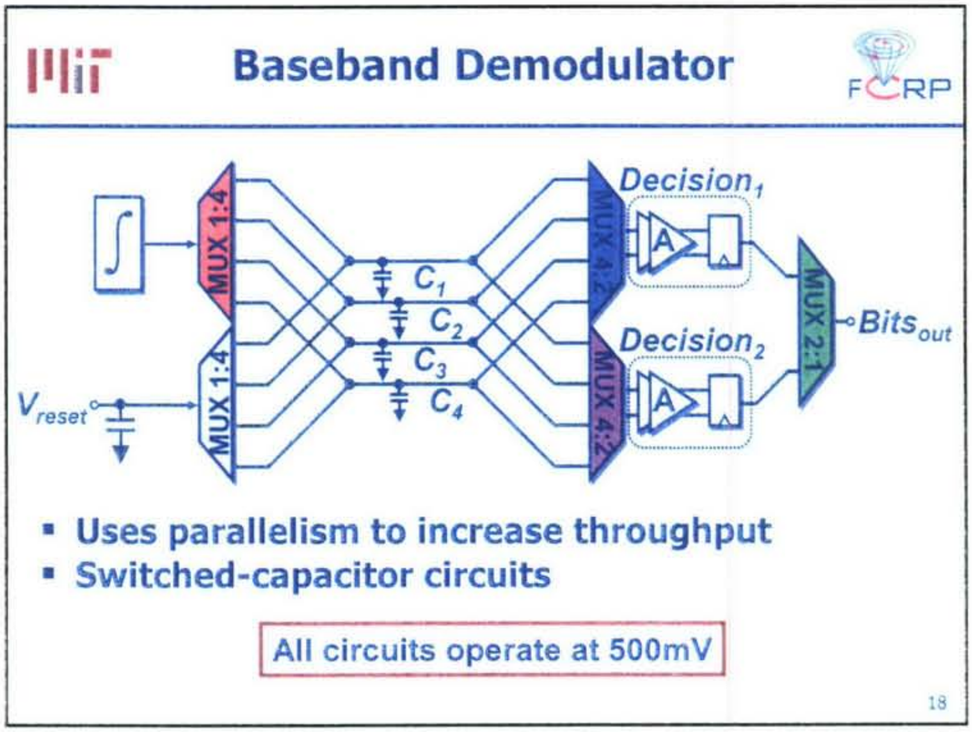
Dynamically biased in 2ns

Single-differential conversion



Passive Self-Mixer

- Fast startup
- Works as a voltage divider
- No bias currents



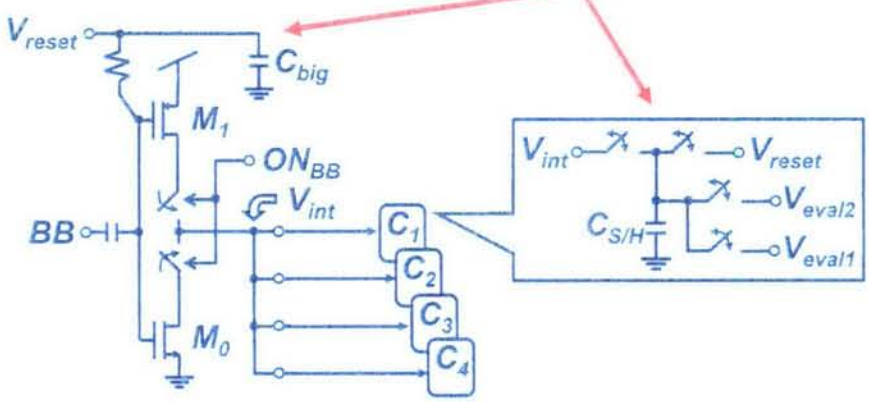
Baseband Demodulator

- Uses parallelism to increase throughput
- Switched-capacitor circuits

All circuits operate at 500mV

0.5V Integrator

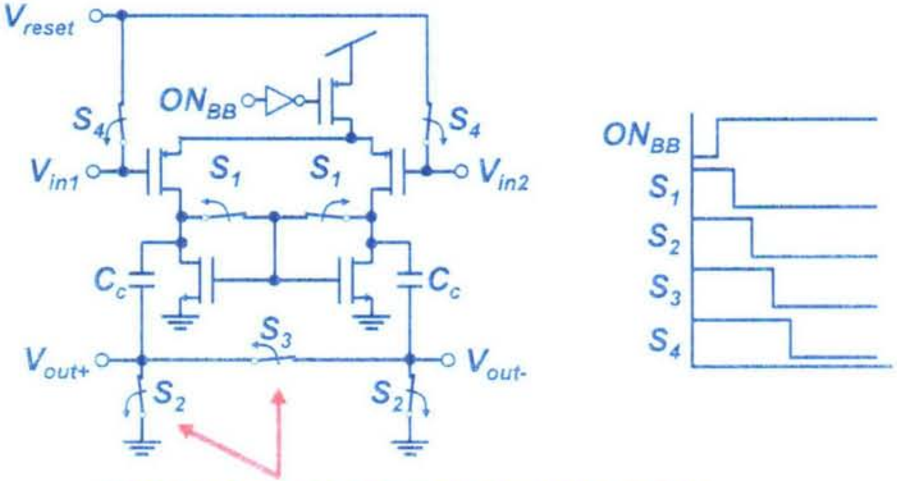
Dynamically biased



Inverter-based integrator

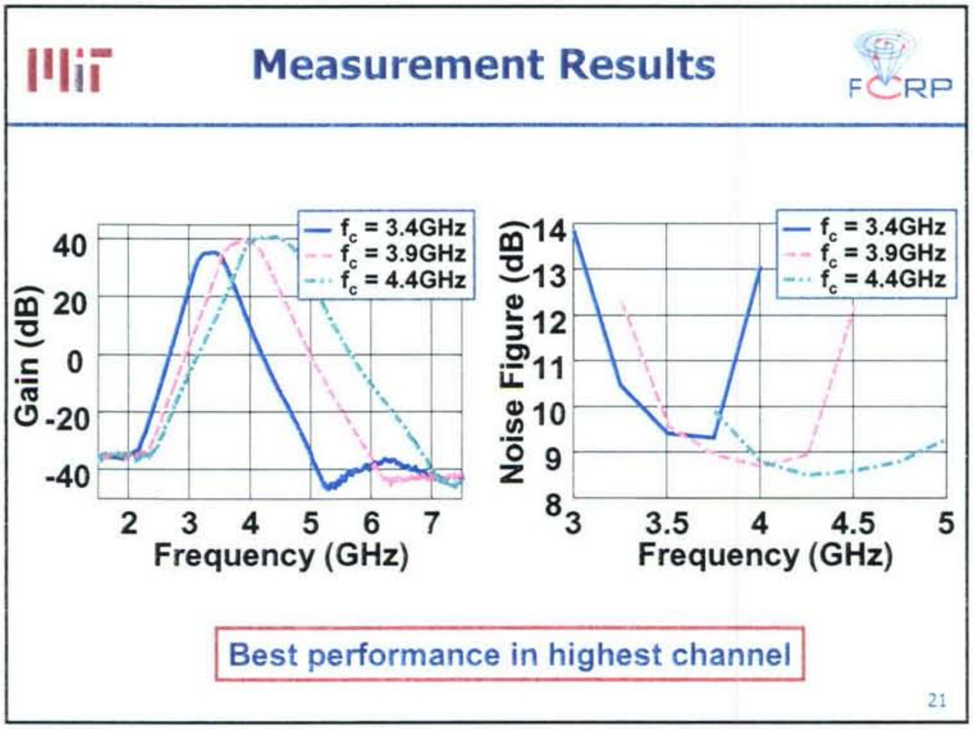
19

0.5V Offset Compensated Preamp



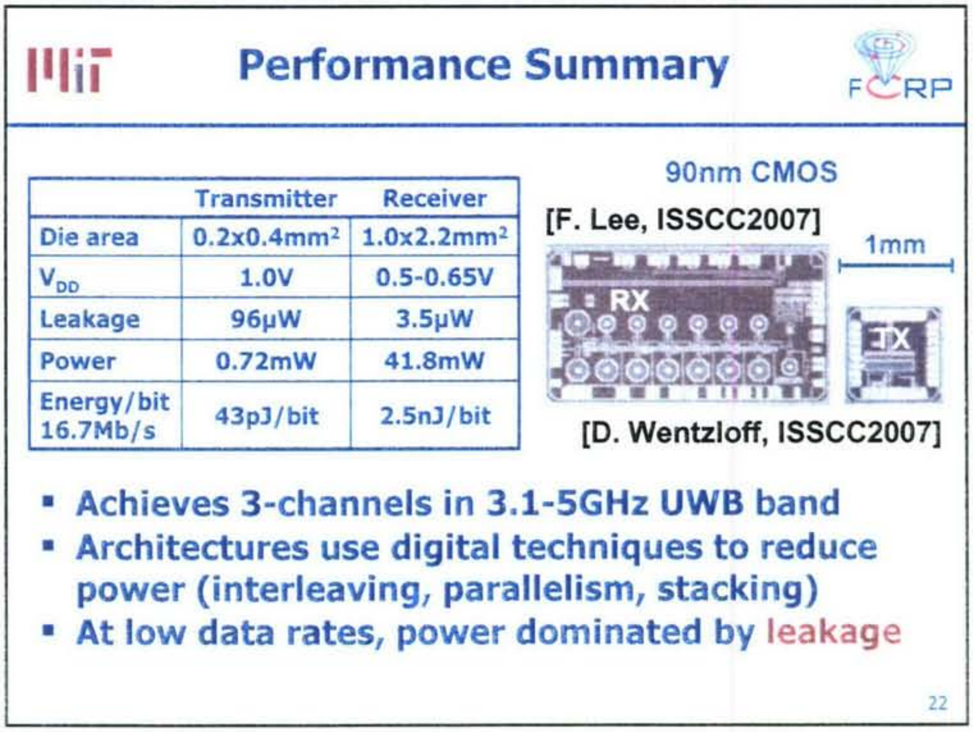
Offset stored on C_c when switches open

20



Measurement Results

Best performance in highest channel



Performance Summary

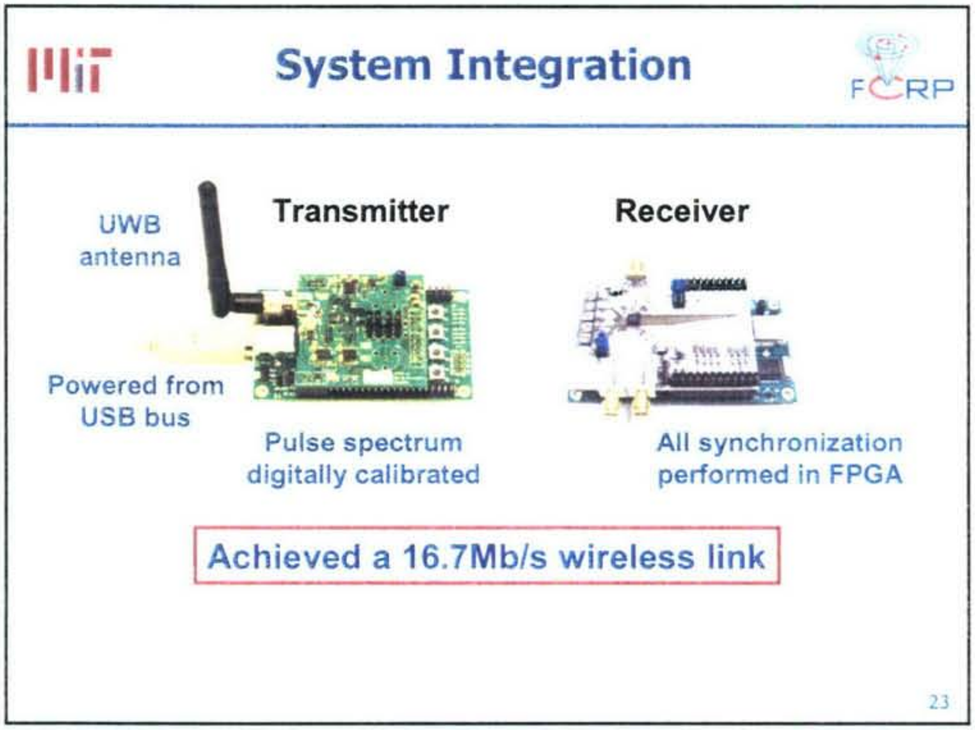
	Transmitter	Receiver
Die area	$0.2x0.4mm^2$	$1.0x2.2mm^2$
V_{DD}	1.0V	0.5-0.65V
Leakage	96µW	3.5µW
Power	0.72mW	41.8mW
Energy/bit 16.7Mb/s	43pJ/bit	2.5nJ/bit

90nm CMOS

[F. Lee, ISSCC2007]

[D. Wentzloff, ISSCC2007]

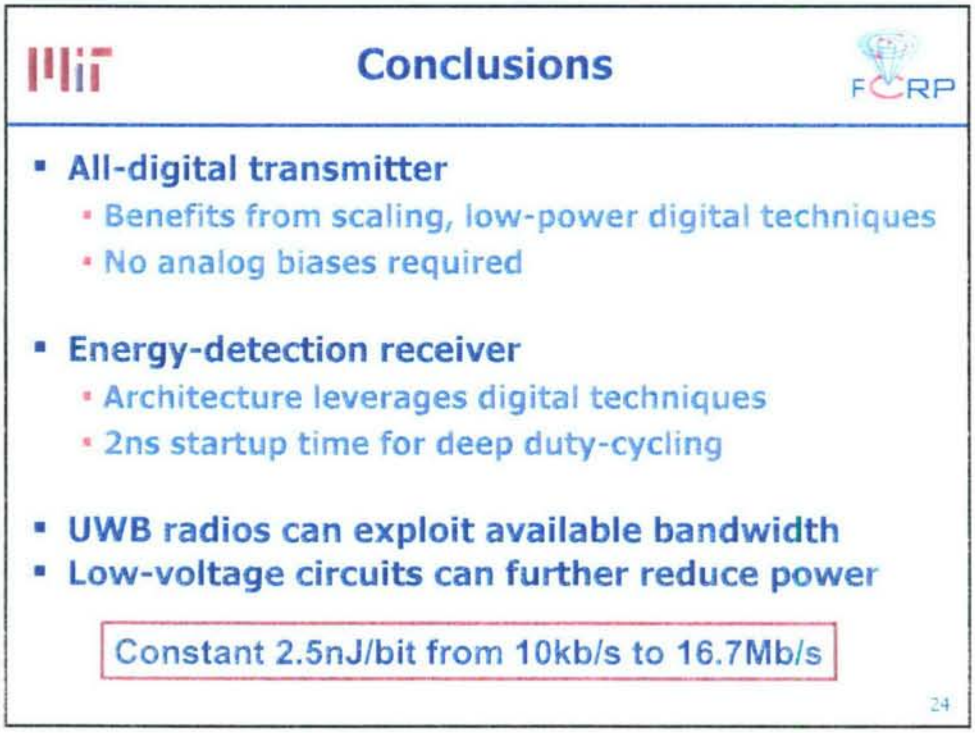
- Achieves 3-channels in 3.1-5GHz UWB band
- Architectures use digital techniques to reduce power (interleaving, parallelism, stacking)
- At low data rates, power dominated by leakage



System Integration

FCRP
UWB antenna
Transmitter
Receiver
Powered from USB bus
Pulse spectrum digitally calibrated
All synchronization performed in FPGA

Achieved a 16.7Mb/s wireless link



Conclusions

- All-digital transmitter
  - Benefits from scaling, low-power digital techniques
  - No analog biases required
- Energy-detection receiver
  - Architecture leverages digital techniques
  - 2ns startup time for deep duty-cycling
- UWB radios can exploit available bandwidth
- Low-voltage circuits can further reduce power

Constant 2.5nJ/bit from 10kb/s to 16.7Mb/s

Recent Conference Publications

34.4 A 256kb Sub-threshold SRAM in 65nm CMOS

Benton H. Calhoun, Anantha Chandrakasan

Massachusetts Institute of Technology, Cambridge, MA

Low-voltage sub-threshold operation has proven to minimize energy per operation for logic [1], and sub-threshold systems will require memories that function at the same low voltages. In this paper, a 65nm SRAM that functions into the sub-threshold region and examines the impact of process variation for low-voltage operation is described.

Previous efforts to reduce SRAM power have included voltage scaling to the edge of sub-threshold [2] or into the sub-threshold region [3], but only for idle cells. Although some published SRAMs operate at the edge of sub-threshold, none function at sub-threshold supply voltages compatible with logic operating at the minimum energy point. The 0.18µm memory in [4] provides one exception. Consisting of latches and using MUX-based read (18T-equivalent bitcell), it operates to 180mV.

Traditional 6T SRAMs face many challenges in deep submicron (DSM) technologies for low V_{DD} operation. Predictions in [5] suggest that process variations will limit standard 90nm SRAMs to around 0.7V operation because of static noise margin (SNM) degradation and write margin, and a V_{DD} of 0.7V is reported for a 65nm SRAM [6]. Measurement results confirm that SNM degradation and inability to write are the two most significant obstacles to sub-threshold SRAM functionality in 65nm. Each of these problems and a bitcell and an architecture that overcomes them, are discussed in this paper.

Figure 34.4.1 shows the impact of local V_T mismatch on the SNM for a standard 6T bitcell in a 65nm process. The Monte-Carlo simulations show that larger channel area decreases the spread of SNM ($\sigma_{VT} \propto 1/\sqrt{WL}$) and that global variation shifts the distribution caused by mismatch [9]. The Hold SNM at 0.3V has roughly the same mean as the Read SNM at 0.5V. However, the 6σ Hold SNM at 0.3V roughly equals the 6σ Read SNM at 0.6V. Likewise, the 6σ Hold SNM at 0.4V and 6σ Read SNM at 0.8V are equivalent. Thus, by eliminating the degraded Read SNM, a bitcell can be operated at 0.3V with the same 6σ stability as a 6T bitcell at 0.6V. A 7T cell avoids Read SNM for above-V_T SRAM [7], but the dynamic storage that it uses is problematic for the longer cycle times of sub-V_T operation.

The 10T bitcell in Fig. 34.4.2 uses transistors M7 to M10 to remove the problem of Read SNM by buffering the stored data during a read access. Thus, the worst-case SNM for this bitcell is the Hold SNM related to M1 to M6, which is the same as the 6T Hold SNM for same-sized M1 to M6. Results from [8] show that single-ended read offers competitive speed for the same area efficiency in DSM. This 10T bitcell uses a full-swing single-ended read that can be 'sensed' using an inverter. Clearly, the extra FETs increase the area by ~66% and also consume leakage power. M10 significantly reduces leakage power relative to the case where it is excluded. In unaccessed cells, M10 prevents node QBB from pulling to '0' even when QB='1'. In this technology, the PMOS sub-threshold current is stronger than NMOS, so node QBB floats close to V_{DD} and decreases sub-threshold current through M8. Also, when QB='0', leakage through M7 is reduced by the stack that M10 creates. Specifically, for iso-V_{DD}, the 10T cell without M10 (a 9T cell) has 50% higher leakage current than the 6T, but adding M10 drops the overhead to 16%. This overhead in leakage current is more than compensated by decreasing V_{DD} by 300mV relative to the 6T bitcell. In simulation, the 10T bitcell at 300mV consumes 2.25× less leakage power than the 6T bitcell at 0.6V (1.75× less relative to 0.5V).

The reduction in sub-threshold leakage through M8 reduces the impact of leakage from unaccessed cells and gives the additional advantage of allowing more cells on a BL during read. Figure 34.4.3 shows the impact of BL leakage on the steady-state voltages while reading a '1' (solid lines) or '0' (dotted lines). For the same number of cells on a BL, the 10T bitcell shows larger BL separation than the 6T (or 9T) bitcells, and 'sensing' with an inverter (whose switching threshold, V_M, is shown) works in simulation from 0°C to 100°C at all corners for 256 cells on a BL. For the 6T cell (or 9T), BL leakage limits the number of cells on a BL to 16 at several process corners for 0.3V. The higher level of integration allowed by the 10T cell reduces the peripheral circuits and slightly mitigates the bitcell area overhead. In order to combat the impact of local V_T mismatch, the WL voltage is boosted relative to the array V_{DD} by 100mV.

Write functionality is the second major obstacle to sub-threshold SRAM, as in this 65nm technology, a 6T bitcell cannot write in the traditional fashion below 0.6V. The plot in Fig. 34.4.4 shows the write margin for the 6T cell under typical and worst-case process corner and temperature. In both cases, the write fails as evident by continued bistability in the cell. Sizing alone cannot correct this problem, because the exponential dependence of sub-threshold drive current on V_T overwhelms the impact of sizing. To achieve write in sub-threshold, the virtual supply (VV_{DD}) to the selected cells floats during the write operation (e.g. [5]). The plot shows that, even for the worst-case, this method provides ample negative noise margin for ensuring a write. Clearly, the side of the bitcell holding a '1' is degraded in voltage due to the collapsing virtual supply. Figure 34.4.4 also shows the essential timing required for the write operation to bring this value to full V_{DD}. The VV_{DD} floats as $\overline{\text{VDDon}}$ is asserted along with WL_WR. The crucial transition in the diagram occurs when $\overline{\text{VDDon}}$ goes low before WL_WR, allowing positive feedback to restore the '1' to full V_{DD}. In the test chip, each row contains a single 128b word that is written at the same time and shares the same VV_{DD}. The block diagram in Fig. 34.4.4 shows how the row is 'folded' so that its cells share a VV_{DD} line.

A 256kb 65nm test chip (Fig. 34.4.7) uses the 10T bitcell and the architecture shown in Fig. 34.4.5. The decoders and other periphery use static CMOS logic for robust sub-threshold operation. The entire array functions at one V_{DD}, and the WL and write drivers operate at 100mV above that supply.

Assuming one redundant row and column are allocated per block, this implementation of the SRAM functions to below 400mV. At 400mV, it consumes 3.28µW and works up to 475kHz. No bit errors for holding data occur in the SRAM until V_{DD} scales below 250mV. Reading works without error at 320mV and writing at 380mV at 27°C. At 85°C, the SRAM writes without error at 350mV and reads without error at 360mV. The measurements on the chip are performed down to 300mV (Fig. 34.4.6 shows correct operation), however at this low voltage mismatch results in bit errors in ~1% of the bits. One type of bit error occurs when a bit holding a '1' is read as a '0' (non-destructive read). This occurs along columns whose I_{RD} has a high V_M due to mismatch. For rows whose M_P is stronger due to mismatch, the write operation fails to overpower M_P sufficiently to flip the contents of the cell, even when VV_{DD} is floating. Both of these problems can be fixed by minor changes to the peripheral circuits, allowing further V_{DD} reduction. Leakage power reduction from V_{DD} scaling is 2.4× and 3.8× relative to 0.6V operation at 0.4V and 0.3V, respectively (Fig. 34.4.6), and active energy savings are 2.25× and 4×.

Acknowledgements:

This work was funded by DARPA and Texas Instruments. We thank Texas Instruments for chip fabrication.

References:

[1] A. Wang, A. Chandrakasan, and S. Kosonocky, "Optimal Supply and Threshold Scaling for Sub-threshold CMOS Circuits," *IEEE Computer Society Annual Symp. on VLSI*, pp. 5-9, Apr., 2002.

[2] N. Kim et al., "Circuit and Microarchitectural Techniques for Reducing Cache Leakage Power," *IEEE Trans. VLSI Systems*, vol. 12, no. 2, pp. 167-184, Feb., 2004.

[3] H. Qin et al., "SRAM Leakage Suppression by Minimizing Standby Supply Voltage," *ISQED*, pp. 55-60, 2004.

[4] A. Wang and A. Chandrakasan, "A 180mV FFT Processor Using Subthreshold Circuit Techniques," *ISSCC Dig. Tech. Papers*, pp. 292-293, Feb., 2004.

[5] M. Yamaoka et al., "Low-Power Embedded SRAM Modules with Expanded Margins for Writing," *ISSCC Dig. Tech. Papers*, pp. 480-481, Feb., 2005.

[6] K. Zhang et al., "A SRAM Design on 65nm CMOS Technology with Integrated Leakage Reduction Scheme," *Symp. VLSI Circuits*, pp. 294-295, June, 2004.

[7] K. Takeda et al., "A Read-Static-Noise-Margin-Free SRAM Cell for Low-Vdd and High-Speed Applications," *ISSCC Dig. Tech. Papers*, pp. 478-479, Feb., 2005.

[8] K. Zhang et al., "The Scaling of Data Sensing Schemes for High Speed Cache Design in Sub-0.18µm Technologies," *Symp. VLSI Circuits*, pp. 226-227, June, 2000.

[9] B. Calhoun and A. Chandrakasan, "Analyzing Static Noise Margin for Sub-threshold SRAM in 65nm CMOS," *ESSCIRC*, pp. 363-366, Sept., 2005.

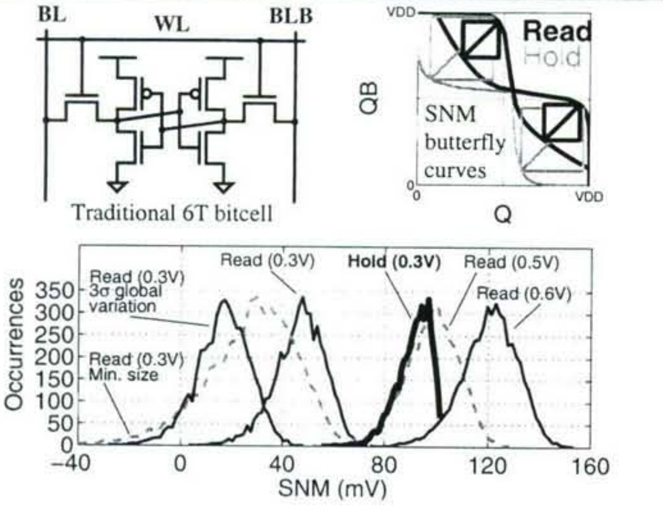


Figure 34.4.1: Impact of local mismatch on 6T SNM in 65nm. Read SNM has larger standard deviation. Hold SNM at 0.3V has roughly the same mean as Read SNM at 0.5V and same 6σ SNM as Read SNM at 0.6V.

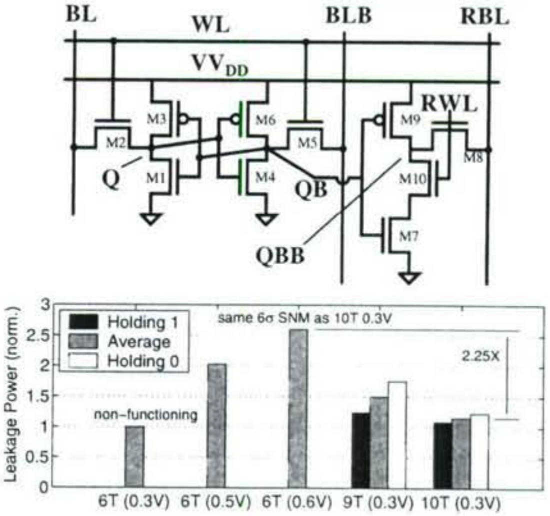


Figure 34.4.2: 10T bitcell for sub-threshold operation. Removing Read SNM allows operation at 0.3V, which leads to 2.25× reduction in leakage power.

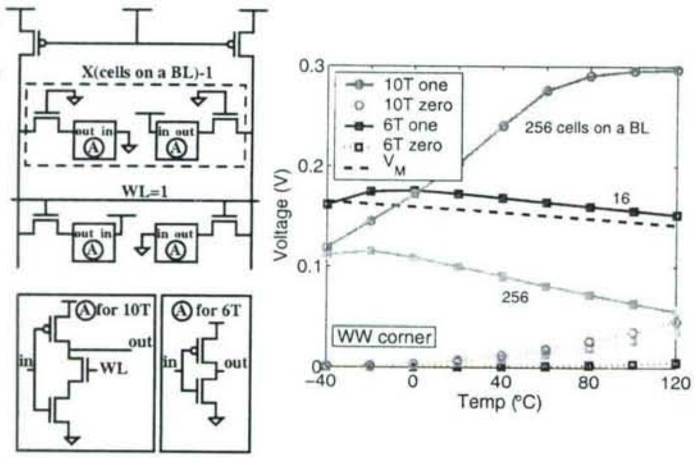


Figure 34.4.3: BL leakage limits the number of cells on a BL. The 10T bitcell can sustain 256 cells/BL at 0.3V compared to 16 without M10 (6T or 9T).

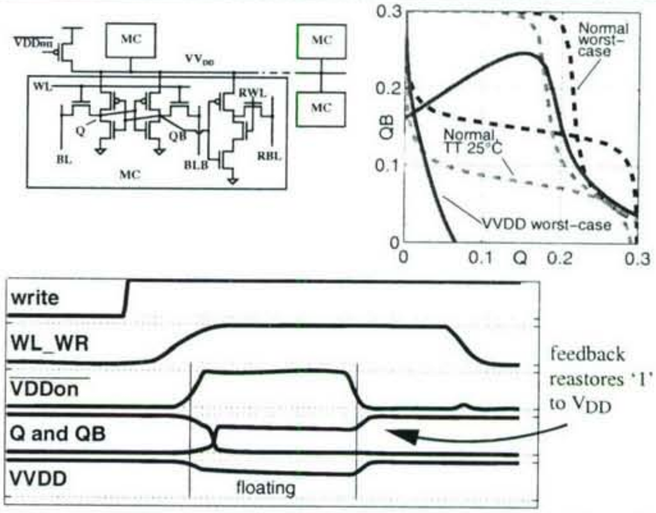


Figure 34.4.4: A virtual supply voltage (VV_{DD}) that floats during write allows robust write operation into sub-V_T (mono-stable butterfly curve). VV_{DD} stops floating while WL_WR remains asserted to restore the '1'value to full V_{DD}.

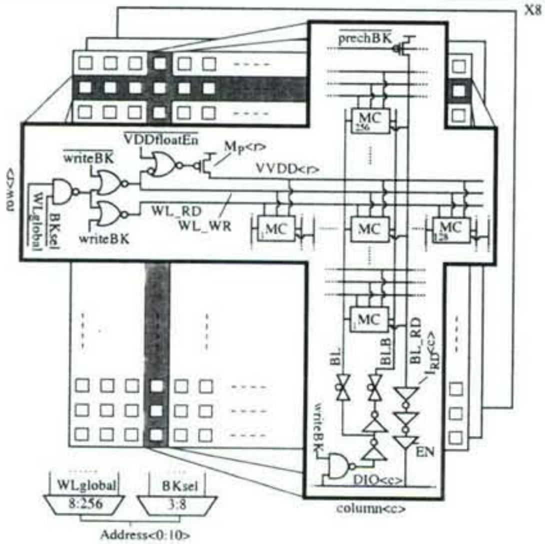


Figure 34.4.5: Architecture of the 256kb test chip.

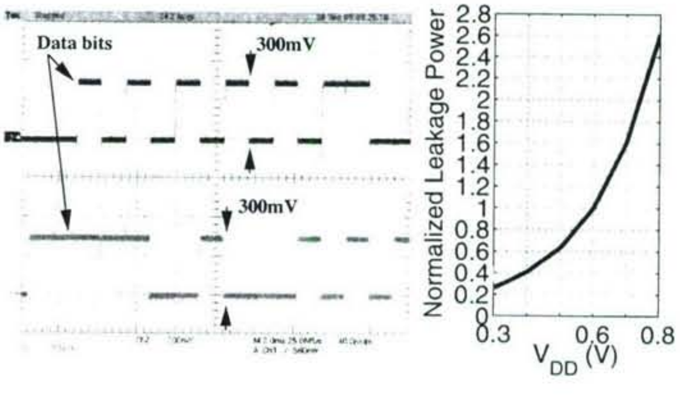


Figure 34.4.6: Chip functioned correctly to below 400mV. Scope plot shows 300mV operation; at this low voltage, some bit errors were observed.

Continued on Page 678

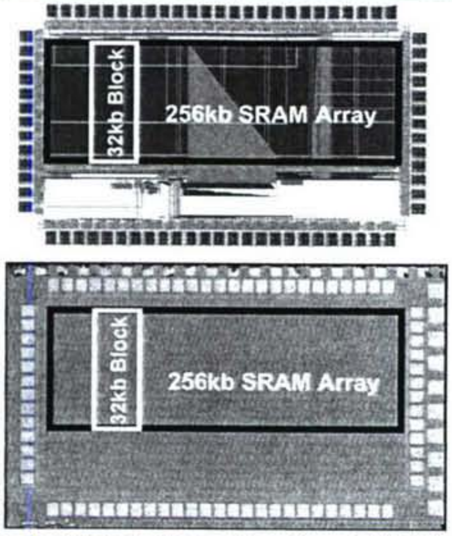


Figure 34.4.7: Annotated die micrograph and layout of 256kb sub-threshold SRAM in 65nm. Die size is 1.88mm × 1.12mm.

AN ENERGY EFFICIENT SUB-THRESHOLD BASEBAND PROCESSOR ARCHITECTURE FOR PULSED ULTRA-WIDEBAND COMMUNICATIONS

V. Sze, R. Blázquez, M. Bhardwaj, A. Chandrakasan
Massachusetts Institute of Technology
sze@mit.edu, rbf@mit.edu, manishb@mit.edu, anantha@mtl.mit.edu

ABSTRACT

This paper describes how parallelism in the digital baseband processor can reduce the energy required to receive ultra-wideband (UWB) packets. The supply voltage of the digital baseband is lowered so that the correlator operates near its minimum energy point resulting in a 68% energy reduction across the entire baseband. This optimum supply voltage occurs below the threshold voltage, placing the circuit in the sub-threshold region. The correlator and the rest of the baseband must be parallelized to maintain throughput at this reduced voltage. While sub-threshold operation is traditionally used for low energy, low frequency applications such as wrist-watches, this paper examines how sub-threshold operation can be applied to low energy, high performance applications. The correlators are further parallelized for a 31x reduction in the synchronization time, which along with duty-cycling, lowers the energy per packet by 43% for a 500 byte packet. Simulation results for a 100Mbps UWB baseband processor are described.

1. INTRODUCTION

The FCC has authorized UWB wireless communications in the 3.1GHz to 10.6GHz band with a minimum bandwidth of 500MHz and a maximum equivalent isotropic radiated power spectral density of -41.3dBm/MHz [1]. IEEE working group 802.15.3a is developing a high data rate standard for wireless personal area networks using UWB.

Applications of UWB include battery-operated devices such as mobile phones, handheld devices and sensor nodes. Consequently, there is a strong demand for an energy efficient UWB system. This paper will describe how operating the digital baseband in the sub-threshold region and increasing the degree of parallelism can translate into energy savings across the entire UWB receiver.

2. UWB SYSTEM ARCHITECTURE

The UWB packets are built from a sequence of binary phase-shift keying pulses with a 500MHz bandwidth. The transmitter generates approximate Gaussian pulses and upconverts the packet to one of 14 channels in the 3.1GHz to 10.6GHz band. Each packet, shown in Figure 1, is divided into two sections: preamble and payload. The preamble contains multiple repetitions of a N_c=31 bit Gold code sent at a pulse repetition frequency (PRF) of 25MHz, or T_{pre}=40ns. The payload contains the actual data and is sent at a PRF of 100MHz, or T_{pay}=10ns, for a 100Mbps data rate with no channel coding.

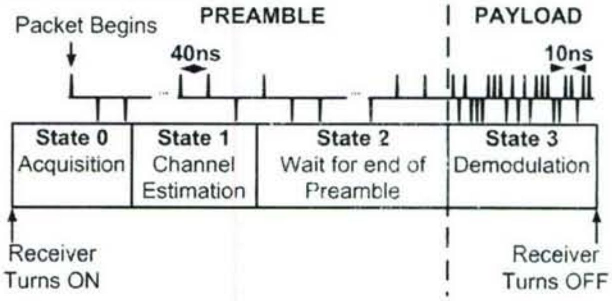


Fig. 1. UWB Packet Format

The receiver, shown in Figure 2, uses a direct conversion architecture in the front-end and the in-phase and quadrature components are sampled at 500MSPS by two 5-bit ADCs. For real-time demodulation of the UWB packet, the digital baseband must perform the signal processing with a throughput of 500MSPS. Synchronization is performed entirely in the digital domain. Only the automatic gain control (AGC) is fed back to the analog domain so that the digital baseband can scale to lower geometries. The baseband was simulated using the digital logic cell library of a 90-nm process.

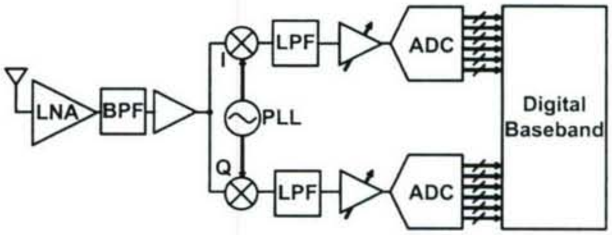


Fig. 2. UWB Receiver

3. DIGITAL BASEBAND PROCESSOR

The digital baseband performs packet detection, acquisition, delay correction and channel estimation using the preamble, followed by demodulation of the payload. Additional repetitions in the preamble are required for AGC, but will not be included in the discussion. Figure 1 outlines the baseband processor's four states of operation with respect to the packet.

In State 0, the acquisition phase, the baseband detects the presence of a packet and provides an initial estimate of its delay. This is accomplished by performing a correlation of the input with an unknown delay against a 31-bit Gold code. Each correlation takes place over $T_{code}=N_c \times T_{pre}=$ 1240ns. The delay must be resolved up to 2ns accuracy; therefore, there are a total of 620 possible delays and corresponding correlations: 20 to match the pulse position, and 31 to match the Gold code. Until acquisition is achieved, the baseband remains in State 0 and performs these correlations. When a correlation exceeds a predefined threshold, acquisition is declared (i.e. lock is detected) and the baseband retimes the input so that it is aligned before moving on to State 1. If all 620 delays are checked and the baseband does not detect lock, the UWB receiver turns off.

In State 1, the channel estimation phase, the baseband must acquire channel estimates from the output of the correlators. This must be done before demodulation in order to compensate for the detrimental effects in the UWB channel [2]. The channel estimates are used to construct a five tap FIR matched filter that takes both the pulse shape and channel impulse response into account.

In State 2, the detection of payload phase, the baseband waits for the end of the preamble which is indicated by an inverted replication of the Gold code. During State 1 and 2, the baseband continuously performs correlations to check that the baseband remains locked. If a threshold is not met, the packet is assumed to be lost or to have been a false packet lock, the baseband and the rest of the UWB receiver turns off. In addition, the baseband performs delay correction with the use of a delay locked loop which is part of the retiming block.

Finally, in State 3, the demodulation phase, each pulse of the payload is filtered by the matched filter derived from the channel estimates and then passed through a decoder that resolves the bit.

A block diagram of the baseband is shown in Figure 3. This paper exploits two forms of parallelism. N defines the degree of parallelism required to operate the digital baseband in sub-threshold. M is defined as the number of Gold Code correlations performed simultaneously. Each sub-bank, composed of N correlators, checks for one Gold Code delay. The trade-offs involved in the specification of M and N will be discussed in the following sections. Other papers have discussed the use of parallelism to reduce *power* consumption for a baseband that uses both autocorrelation and cross-correlation [3]; however, the metric here is to reduce the *energy* consumption for a baseband that uses only cross-correlation.

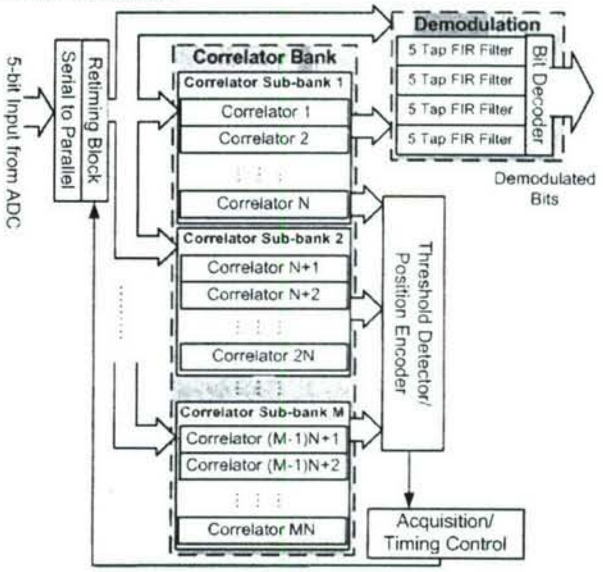


Fig. 3. UWB Parallelized Digital Baseband

4. SUB-THRESHOLD OPERATION (IMPACT OF N)

As previously mentioned, since the input from the ADC arrives at a rate of 500MSPS, a serial baseband must run at a frequency of 500MHz if the input is to be processed in real time. In order that the critical paths, through the correlator and through the matched filter, meet the timing constraint, the digital circuitry must run at its maximum supply voltage. However, running at the maximum voltage is not energy-efficient. It is important to reduce the energy of the correlator since it consumes the largest portion of energy in the baseband during synchronization. The energy per operation can be reduced by lowering the supply voltage (V_{dd}) [4]. At maximum V_{dd}, the transistors in the circuit operate in the active region. If V_{dd} is lowered below the threshold voltage (V_{th}) of the device, the circuit is said to be operating in the sub-threshold region. Lowering V_{dd} increases the latency per operation (T_{period}) linearly in the active region, and exponentially in the sub-threshold region. This increases the leakage energy as it is linearly related to T_{period}. There is a minimum operating energy point since the dynamic energy and the leakage energy scale in an opposite manner with V_{dd} [5]. Spectre simulations of the correlator in the 90-nm process show that operating at the minimum energy point of 0.3V rather than at the maximum V_{dd} of 1V reduces the energy per operation of the correlator by 89% (Figure 4).

At the minimum energy point, the baseband processing must be parallelized to maintain a throughput of 500MSPS. For ease of design, it is desirable that the PRF of the

1-4244-0469-X/06/$20.00 ©2006 IEEE

preamble be a multiple of the clock frequency of the baseband. Since this is not possible at 0.3V, the baseband operates slightly above the minimum energy point at 0.4V with a frequency of 25MHz which requires N=20 correlators to form a sub-bank of correlators. Lowering the supply voltage from 1V to 0.4V (sub-threshold) results in an overall energy savings of 83% for the correlators and 68% for the entire baseband. The energy savings for the entire baseband is less since buffers are inserted in some paths to compensate for the increased transition time at 0.4V.

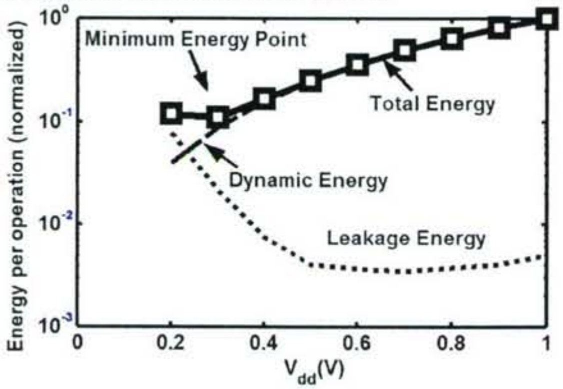


Fig. 4. Simulated energy plot for the correlator

5. REDUCED ACQUISITION TIME (IMPACT OF M)

Increasing M increases the number of code shifts that are simultaneously checked. Although this increases the power consumed by the baseband during acquisition, the time spent in acquisition decreases proportionally. In this section a model is developed to show that the baseband energy remains approximately the same for any M, while the energy spent by the rest of the receiver scales inversely with M. This results in an overall reduction in energy per packet.

5.1. Modeling Energy per Packet.

The average time and amount of energy the baseband spends in each state must be determined. The total time the baseband spends in State 0 and 2 is set by the number of times the Gold code is repeated in the preamble, R(M), which can be reduced by increasing parallelism M.

$$R(M)=\left\lceil \frac{N_c}{M} \right\rceil \tag{1}$$

While the time spent in State 1 and State 3 is fixed, the distribution of time between State 0 and 2 is dictated by when the baseband detects lock. The maximum time the baseband will remain in State 0 is R(M)×T_{code}. In this case, no time is spent in State 2.

Let D be the number of code shifts between the Gold code in the preamble and the Gold code in the baseband. Assume that D is uniformly distributed over [0, N_c-1]. Let I(D,M) be the number of code durations (T_{code}) required to achieve acquisition in State 0. Let P_d be the probability that the baseband detects lock when the input and the Gold code are aligned, and $P_{fa\text{-}M}$ be the probability that the baseband detects lock in one or more of M delays which are not aligned to the code. For small P_{fa} (=$P_{fa\text{-}1}$), $P_{fa\text{-}M} \approx$ M×P_{fa}.

Assuming the detector is ideal (P_d=1, $P_{fa\text{-}M}$=0),

$$I(D,M)=\left\lceil \frac{D}{M} \right\rceil \tag{2}$$

The maximum number of code durations, T_{code}, required to achieve acquisition in State 0 is R(M). As the baseband performs different operations during each state, the energy per T_{code} varies per state. E_0 and E_2 are the energies consumed over T_{code}, in State 0 and State 2, while E_1 and E_3 are the energies required to perform channel estimation and demodulation respectively. It is important to note that E_0, to a first order, scales linearly with M. After acquisition, M-1 of the correlator sub-banks can be turned off through the use of clock gating and power gating so that E_1, E_2 and E_3 are not dependent on M to a first order.

The energy per packet is computed as follows,

$$\begin{aligned} Energy(D,M) &= \sum_{X=1}^{R(M)} \Pr(X,D,M)\times Energy(X,D,M) \\ &= \alpha E_0+\beta E_1+\gamma E_2+\varepsilon E_3 \end{aligned} \tag{3}$$

Pr(X,D,M) is the probability that the baseband will stay in acquisition for X units of T_{code} given a packet with delay D. Energy(X,D,M) is the energy consumed by the baseband.

For all X≠I(D,M),

$$\Pr(X,D,M)=P_{fa-M}\left(1-P_{fa-M}\right)^{X-1} \tag{4}$$

$$Energy(X,D,M)=\left(XE_0+E_1\right) \tag{5}$$

When X=I(D,M),

$$\begin{aligned} &\Pr(X,D,M)\times Energy(X,D,M) \\ &=P_d\left(1-P_{fa-M}\right)^{X-1}\left\{XE_0+E_1+\left(R(M)-X\right)E_2+E_3\right\} \\ &\qquad +(1-P_d)\left(1-P_{fa-M}\right)^{X-1}R(M)E_0 \end{aligned} \tag{6}$$

This paper assumes P_d=0.9 and P_{fa}=10^{-5}, which were derived from the 802.15.3a proposal. The average energy required by the baseband to process a packet for a given degree of parallelism M is computed by taking the expected value of the energy per packet over all possible delays D, conditioned on M. If P_{fa} is small, this average baseband energy does not change significantly since, to a first order, the same number of operations occur for any M. In addition, for a small P_{fa}, the required preamble time and hence energy spent during acquisition by the rest of the receiver scale inversely with M.

5.2. Impact of M on Energy per Packet

The energy per packet can be broken down into the preamble energy and the payload energy. The payload energy is fixed by the number of bits transmitted per packet. However, the length of the preamble, and consequently the preamble energy, can be reduced based on the configuration of the baseband. A previous version of the UWB baseband checked one combination of the Gold code at a time [6]. In order to check all shifted combinations of the 31-bit Gold code, the baseband must perform at least 31 correlations. The preamble must last for $N_c \times T_{pre} \times R(M=1) = 34.880\mu s$.

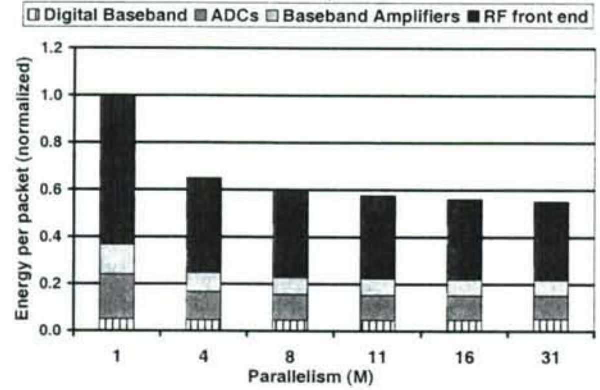


Fig. 5. Average packet energy consumption of the receiver subsystems for various degrees of parallelism.

By using multiple sub-banks of correlators that operate in parallel, the number of Gold code shifts that can be checked in one cycle is increased, which reduces the number of repetitions required in the preamble. In a fully parallelized baseband, with 31 sub-banks of correlators, all 31 shifted possibilities of the Gold code are checked simultaneously, and the Gold code only has to be repeated once in the preamble for acquisition. This results in a 31x reduction in the preamble length. As previously stated, for varying degrees of M, the energy spent by the baseband on the acquisition is almost the same with a slight increase due to increased interconnect capacitance that results from parallelism. The actual energy savings result from the other circuitry in the UWB receiver. Reduction in acquisition time implies that the entire receiver needs to be on for a much shorter period of time. The RF front end, ADCs and the baseband amplifiers can be turned off once the packet has been demodulated. The measured power of these blocks is approximately 79% of the receiver power [7], [8]; shutting them off earlier translates into significant energy savings. Figure 5 shows the reduction in energy per packet, with payload size of 500 bytes, for various degrees of parallelism. It can be concluded that faster synchronization, combined with duty-cycling, reduces the energy required to receive a UWB packet. As increasing M only affects preamble energy, the impact of using parallelism to reduce energy per packet varies with payload size (Figure 6).

It is important to note that reduction in preamble energy should not be made at the expense of the payload energy. Techniques such as clock gating and power gating are used to ensure that the power consumption of the baseband during demodulation does not increase with parallelization. During State 1, 2 and 3, either most or all correlators are turned off and power gated to reduce leakage, and clock gating should be used to reduce the impact of the increased interconnect capacitance due to parallelism.

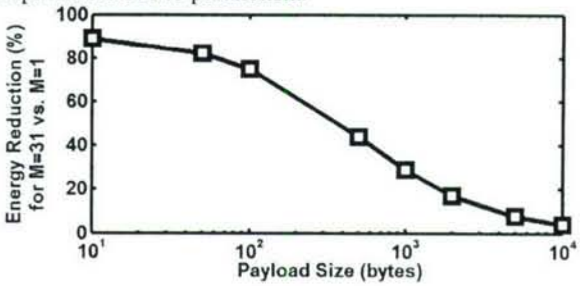


Fig 6. Energy reduction for various payload sizes

6. CONCLUSION

This paper discusses how parallelism allows for voltage scaling and reduced acquisition time, which reduces the energy required to receive a UWB packet. Voltage scaling to sub-threshold allows the correlator sub-banks to operate near the minimum energy point, resulting in an energy per operation reduction in the correlators of 83% and energy reduction of 68% across the entire baseband. The reduced acquisition time through further parallelization of the correlator sub-banks by 31 led to a 43% reduction in energy per packet for a 500 byte packet. The analysis in this paper can be mapped to other high performance communication applications using sub-threshold operation and parallelism.

ACKNOWLEDGMENTS

This research is sponsored by DARPA and an NSERC Fellowship. We thank STMicroelectronics for providing process and cell library access.

REFERENCES

[1] FCC, "Ultra-wideband First Report and Order", February 2002.
[2] J. Foerster, "Channel Modeling Sub-Committee Report Final," IEEE P802.15 Working Group for Wireless Wireless Personal Area Networks (WPANs). February 2002.
[3] M. Verhelst, W. Dehaene, "System Design of an Ultra-low Power, Low Data Rate, Pulse UWB Receiver in the 0-960MHz Band," *IEEE International Conference on Communications*, 2005.
[4] A. Wang, A. Chandrakasan, "A 180mV FFT Processor Using Subthreshold Circuit Techniques," *ISSCC*, February 2004.
[5] B. H. Calhoun, A. P. Chandrakasan, "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits," *ISLPED*, August 2004.
[6] R. Blázquez, A. Chandrakasan, "Architectures For Energy-Aware Impulse UWB Communications," *ICASSP*, 2005.
[7] B. P. Ginsburg, A. Chandrakasan, "Dual Scalable 500MS/s, 5b Time-Interleaved SAR ADCs for UWB Applications,"*CICC*, 2005.
[8] F. S. Lee , A. P. Chandrakasan, "A BiCMOS Ultra-Wideband 3.1-10.6GHz Front-End," *CICC*, 2005.

A 500MS/s 5b ADC in 65nm CMOS

Brian P. Ginsburg and Anantha P. Chandrakasan

Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology
50 Vassar Street, Room 38-107
Cambridge, MA 02139, USA
Tel: 617-258-6405, Fax: 617-253-5053, email: bginzz@mit.edu

Abstract

A 1.2V 6mW 500MS/s 5-bit ADC for use in a UWB receiver has been fabricated in a pure digital 65nm CMOS technology. The ADC uses a 6-channel time-interleaved successive approximation register architecture. Each of the channels has a split capacitor array to reduce switching energy and sensitivity to digital timing skew. A variable delay line is used to optimize the instant of latch strobing to reduce preamplifier currents.
Keywords: UWB, ADC, SAR, CMOS, time-interleaved

Introduction

Ultra-wideband (UWB) radio is an emerging technology that shows promise for very-high-data-rate wireless communication over short distances. High speed (>500MS/s) and low resolution (4-5b) ADCs are required to convert these signals. It is desirable for integration of the ADC directly with the high-performance UWB digital baseband processor in a deep sub-micron CMOS process for best digital performance. Low-power time-interleaved successive approximation register (SAR) ADCs have been demonstrated at the speeds necessary for UWB radio [1], [2]. The SAR topology is well suited for implementation in deep sub-micron CMOS due to its very low analog complexity.

This paper presents a 500MS/s 5-bit ADC in pure-digital 65nm CMOS. The ADC has 6 time-interleaved channels synchronized to a common clock; each channel uses six clock periods to perform a conversion (one for sampling followed by five bit-cycles); thus the channels sample sequentially every clock period. The ADCs have been designed to take advantage of the process technology without sacrificing robustness in the presence of increased variability. Two new techniques are incorporated to improve energy-efficiency. A split capacitor array reduces switching energy and is robust to digital delay mismatches. In the comparator, a variable delay line and on-chip delay detector optimize the instant of strobing for the regenerative latch to lengthen settling times for preamplifiers.

Technology Considerations

Deep sub-micron CMOS provides both opportunities and challenges for mixed-signal design. The SAR architecture can benefit greatly from reduced features sizes because it has significant digital but little analog complexity. The two principal analog blocks in a SAR converter are the capacitor array DAC and the comparator. The former benefits directly from the reduced gate length and lower on resistance of the switches. Sampling at the lower power supply (1.2V) is achieved by constraining the input voltage to the 0-0.4V range; thus a standard V_T NMOS samples the input. The comparators use a two stage preamplifier and a regenerative latch. Each preamplifier, seen in Fig. 2(a), uses non-minimum length input transistors to improve both matching and output impedance. While this increases the device capacitance for the same g_m, the presence of wiring parasitics reduces the overall impact.

Split Capacitor Array

The DAC is the first implementation of the split capacitor array, wherein the MSB capacitor is split into an identical copy of the rest of the array, theoretically analyzed in [3]. This array is predicted to have 37% lower switching energy than the conventional array and 1-step switching method without any increase in total capacitance or area. Besides the energy savings presented in [3], the split capacitor array is also well suited for high-speed implementations. In a conventional array, when two capacitors are required to transition on a given bit-cycle, variation in digital propagation delays can cause the array output to initially transition in the wrong direction, producing a large overdrive condition for the preamplifiers, increasing their settling times. In the split capacitor array, only one capacitor switches during any bit-cycle, providing inherent robustness against these digital timing skews, as shown in Fig. 1. Under the worst-case timing skew, the settling time is reduced by 10%.

Optimized Latch Strobing

During bit-cycling, the clock period is divided into one phase for the settling of the DAC and preamplifiers and one phase for regeneration of the latch. The latch typically resolves in much less than one 1ns even for very small inputs. The ADC sits idle after the latch settles until the start of the next bit-cycle. Self-timed bit-cycling has been proposed to use this idle time to start the next bit-cycle early [4]. This approach relaxes the preamplifier settling time requirement for all but the first bit-cycle (determining the MSB), as it has no prior bit-cycle from which to borrow. Here, a variable delay line has been inserted in series with the latch strobe signal (Fig. 2(a)) to extend analog settling time in the first half of every bit-cycle, including the first, "pre-borrowing" time from that bit-cycle's own latch phase. The slower speed requirements

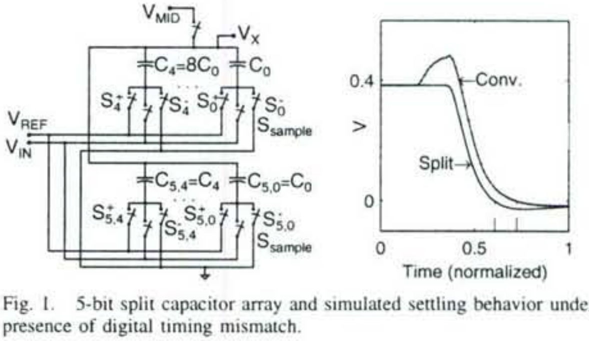


Fig. 1. 5-bit split capacitor array and simulated settling behavior under the presence of digital timing mismatch.

allows reduced preamplifier currents. To tune this delay for various clock frequencies and operating conditions, an on-chip delay detector has been designed, shown in Fig. 2(b). The latch's inputs are shorted to produce the worst case settling behavior and its outputs are captured both by a replica of the SAR digital path (R_1–R_2) and the *Done* signal in R_3–R_4. Any difference between these outputs is an indication of the failure of the latch to resolve fast enough to meet the setup time constraints of R_1–R_2, and thus the delay should be reduced. An off-chip loop is used to determine the frequency of errors and tune the delay via a configuration register. This function could be implemented on-chip with a counter and a simple finite state machine.

Measured Results

The ADC has been fabricated in a pure-digital 65nm CMOS technology with a nominal supply voltage of 1.2V. At 500MS/s, the analog and digital supplies, excluding I/O power, consume 2.86mW and 3.06mW, respectively. Using a separate on-chip test channel with the conventional array and switching method, the measured DAC energy savings for the split capacitor array is 31%, which closely matches the theoretical model; increased bottom-plate routing accounts for the difference. The static linearity is $-0.16/0.15$ INL and $-0.25/0.26$ DNL (Fig. 3); the split capacitor array shows no linearity degradation versus the conventional array. The dynamic results are presented in Fig. 4. The SNDR does not drop by 3dB until past the Nyquist frequency. An FFT of a 239.04 MHz input sampled at 500MS/s is shown in Fig. 5. The level of offset voltage mismatch and timing skew is sufficiently low for proper reception of UWB signals. Using the figure of merit in [1], $(P/(2^{ENOB}2f_{in}))$, at the Nyquist frequency, the ADC achieves 755fJ/conv. step. At 250MS/s, 420fJ/conv. step is achieved by lowering the voltage supplies. The die photograph is shown in Fig. 6.

Acknowledgments The authors would like to thank Texas Instrumnts for fabricating the chip. This work is funded by NSF, DARPA, and an NDSEG Fellowship.

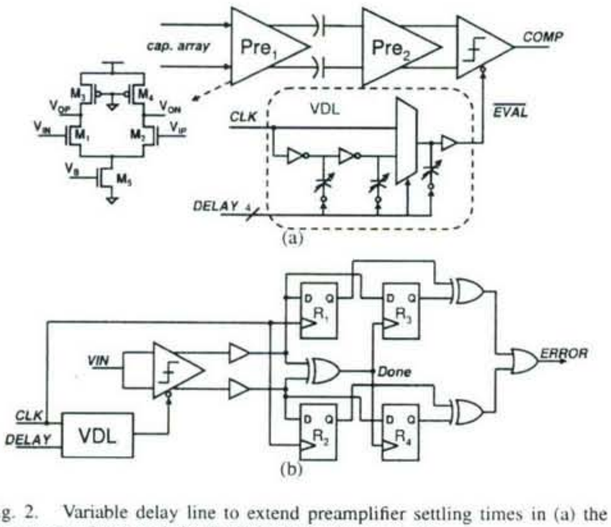


Fig. 2. Variable delay line to extend preamplifier settling times in (a) the comparator circuit and (b) the latch-delay-detect circuit.

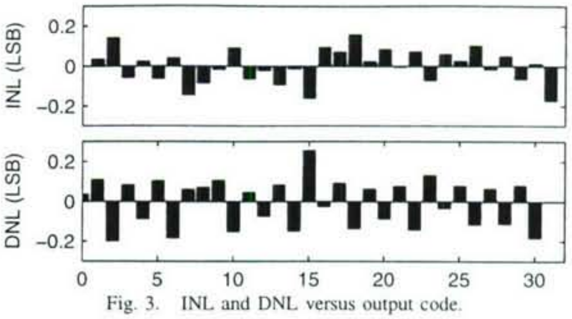


Fig. 3. INL and DNL versus output code.

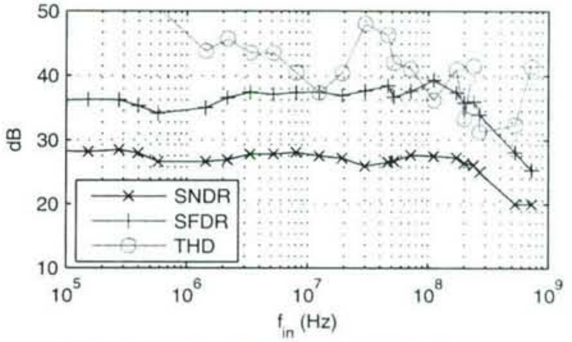


Fig. 4. SNDR and SFDR versus input frequency.

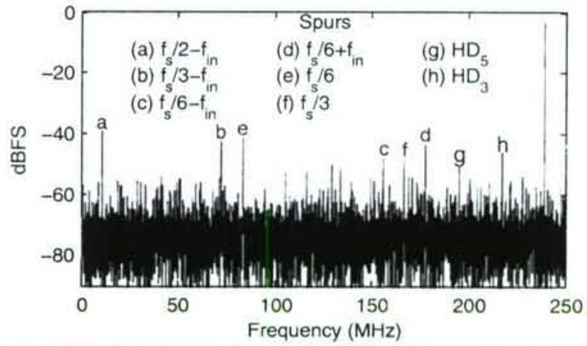


Fig. 5. FFT of 239.04MHz sine wave sampled at 500MS/s with dominant spurs labeled. (a)-(d) are from timing skew, and (e)-(f) are from offset mismatch

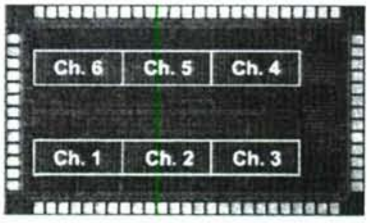


Fig. 6. Photograph of 1.9 × 1.2mm die.

References

[1] D. Draxelmayr, "A 6b 600MHz 10mW ADC array in digital 90nm CMOS," in *ISSCC Dig. Tech. Papers*, Feb. 2004, pp. 264–265.

[2] B. P. Ginsburg and A. P. Chandrakasan, "Dual scalable 500MS/s, 5b time-interleaved SAR ADCs for UWB applications," in *Proc. of the IEEE 2005 Custom Integrated Circuits Conference*, Sept. 2005, pp. 10.7.1–10.7.4.

[3] B. P. Ginsburg and A. P. Chandrakasan, "An energy-efficient charge recycling approach for a SAR converter with capacitive DAC," in *Proc. of the IEEE Int. Symp. on Circuits and Systems*, May 2005, pp. 184–185.

[4] G. Promitzer, "12-bit low-power fully differential switched capacitor noncalibrating successive approximation ADC with 1MS/s," *IEEE J. Solid-State Circuits*, vol. 36, no. 7, pp. 1138–1143, July 2001.

Variation-Driven Device Sizing for Minimum Energy Sub-threshold Circuits

Joyce Kwong
jyskwong@mtl.mit.edu

Anantha P. Chandrakasan
anantha@mtl.mit.edu

Massachusetts Institute of Technology
50 Vassar St., Room 38-107
Cambridge MA 02139, USA

ABSTRACT

Sub-threshold operation is a compelling approach for energy-constrained applications, but increased sensitivity to variation must be mitigated. We explore variability metrics and the variation sensitivity of stacked device topologies. We show that upsizing is necessary to achieve robustness at reduced voltages and propose a design methodology to meet yield constraints. The need for upsizing imposes an energy overhead, influencing the optimal supply voltage to minimize energy. Finally, we characterize performance variability by summing delay distributions of each stage in an arbitrary critical path and achieve results accurate to within 10% of Monte Carlo simulation.

Categories and Subject Descriptors: B.8.1 [Reliability, Testing, and Fault-Tolerance]

General Terms: Performance, Design, Reliability

Keywords: Sub-threshold circuits, Minimum energy point, Delay model

1. INTRODUCTION

In sub-threshold circuits, the power supply is set below the transistor threshold voltage V_T to obtain energy savings when speed is not the primary constraint [1]. Authors of [2][3] derived analytical expressions for the optimum V_{DD} to minimize energy in sub-threshold and showed its dependence on major circuit parameters. Sub-threshold circuits rely on leakage currents that are exponentially dependent on V_T and are therefore more sensitive to process variation than traditional above-threshold designs.

It was suggested in [4] that minimum size devices are theoretically optimal for minimizing energy in sub-threshold. However, minimum size devices have increased sensitivity to V_T variation because σ_{V_T} is roughly proportional to $(WL)^{-\frac{1}{2}}$. If a minimum size circuit does not function at the optimum V_{DD} due to degraded logic output swing, it is necessary to upsize devices to improve robustness at the expense of increased energy consumption. Therefore, variability must be considered when analyzing the minimum energy operating point.

Previous work in [5] addresses intra-die variation by providing statistical models for energy and delay of an inverter chain in sub-threshold. An empirical expression for the optimum voltage is shown as a function of logic depth, assuming complete functionality at V_{min}. Work in [6] presents a unified delay variability expression for strong- and weak-inversion and applies it to a NAND gate. Researchers have also proposed various approaches to optimize delay yield by tuning V_{DD}/V_T or choosing gates of different drive strengths, for example in [7]. However, functional yield was not considered until [8][9], which address unsatisfactory V_{OH} and V_{OL} in sub-threshold inverters whose output levels are degraded by leaking devices, such as in a register file. Body biasing is another option for mitigating variation in sub-threshold [10] when a triple-well process is available.

We address inter- and intra-die variation and show that functionality in sub-threshold circuits may be compromised without proper design for variations. We first explore variability metrics for the inverter and logic gates with stacked devices, and propose a metric to size logic gates for a fixed failure rate under process variation. We then examine the energy versus V_{DD} profile given the failure rate constraint and find the optimum sizing and supply voltage. We present an efficient methodology to model delay variability of a chain of logic gates and characterize the effect of yield-based sizing constraints on performance variability.

2. VARIABILITY METRICS AND DEVICE SIZING

A commonly used expression for sub-threshold current is given by [11]

$$I_{sub} = I_o e^{\frac{V_{GS}-V_T+\eta V_{DS}}{nV_{th}}}(1-e^{\frac{-V_{DS}}{V_{th}}}) \qquad (1)$$

$$I_o = \mu_o C_{ox}\frac{W}{L}(n-1)V_{th}^2 \qquad (2)$$

where n is the sub-threshold swing factor, V_{th} the thermal voltage, and η the DIBL coefficient. The nominal current scales linearly with W/L, while standard deviation of V_T distribution reduces with $(WL)^{-\frac{1}{2}}$, thus lowering sub-threshold current variation. This section explores how sizing affects

ISLPED'06, October 4–6 2006, Tegernsee, Germany
Copyright 2006 ACM 1-59593-462-6/06/0010 ...$5.00.

variability in output swing and active current in the inverter and stacked device topologies.

2.1 Logic Gate Output Swing

In the sub-threshold regime, the ratio of active to idle currents in a logic gate is much lower than in strong inversion. If, for example, process variation strengthens NMOS relative to PMOS, a pull-up network will not be able to drive the logic gate output fully to V_{DD} because of idle leakage in the pull-down network. This degradation in gate output swing is illustrated in Figure 1(a). The solid line shows the voltage transfer characteristic (VTC) of a minimum size inverter in a 65nm technology at skewed global process corner. Dashed lines plot the VTCs when random local V_T mismatch is applied to the inverter. One case shows a severely degraded V_{OL}, which can cause functional error if it is above the input low threshold (V_{IL}) of the succeeding gate. Therefore, V_T variation significantly impacts circuit functionality in deeply scaled technologies.

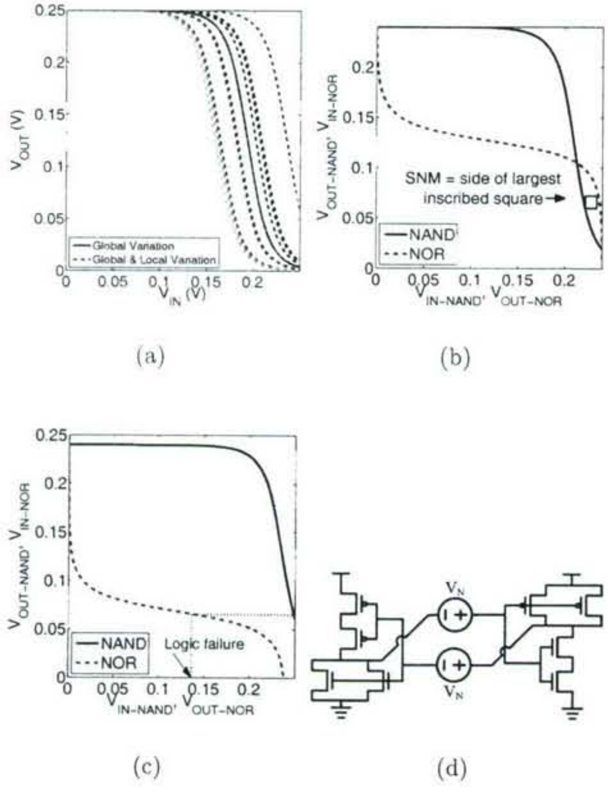


Figure 1: (a) Inverter VTCs at skewed process corner with random V_T mismatch. (b) Butterfly plot of NAND/NOR gates with functional output levels. (c) Butterfly plot of NAND with failing V_{OL}. (d) Example circuit for verifying logic gate output levels.

A consistent metric is necessary to determine whether a logic gate has sufficient V_{OL} and V_{OH} levels. Arbitrary limits, such as 10% and 90% of V_{DD}, do not scale well across global process corners. For example, at the strong-PMOS weak-NMOS corner, strong leakage through PMOS raises V_{OL} of all gates above ground. This also shifts VTCs to the right, and thus logic gates can tolerate higher V_{OL} in the preceding gate. Instead of arbitrary limits, we propose using butterfly plots to verify output voltage levels, specifically in the context of standard cell design.

2.1.1 *Use of the Butterfly Plot*

To verify V_{OL} of a given gate, we superimpose its VTC with the mirrored VTC of NOR, since the latter has the most stringent V_{IL} requirement from stacked devices in the pull-up network and parallel devices in the pull-down. Similarly, we verify V_{OH} using the NAND VTC, which has the worst case V_{IH}.

In Figure 1(b), a NAND gate has sufficient output swing such that $V_{OL-NAND}$ produces a logic high output in a succeeding NOR gate. In contrast, the NAND gate in Figure 1(c) exhibits $V_{OL-NAND}$=65mV and produces a NOR output of 136mV, close to mid-rail and thus causing logic failure.

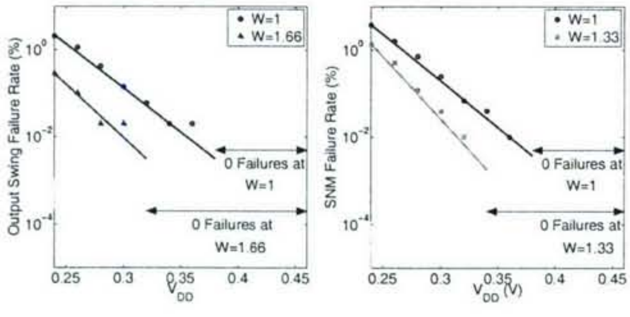
A gate with failing output levels is analogous to a 6T SRAM cell displaying negative static noise margin (SNM), in that the butterfly plots for both cases do not contain an inscribed square. Therefore, we can also apply [12] to find the side of the largest inscribed square, illustrated in Figure 1(b). Figure 1(d) shows an equivalent circuit for this measurement on two back-to-back logic gates. Because the VTC is input-dependent, all inputs are varied simultaneously to obtain the worst case V_{IH} and V_{IL}.

It was shown in [13] that the SNM of two back-to-back gates G1 and G2 is equal to the maximum noise that can be applied to all gates in an infinitely long chain of alternating G1 and G2, before logic failure occurs. Thus when verifying a standard cell G using the butterfly plot, we essentially assume that all logic paths in a synthesized circuit are composed of alternating G and NAND3 gates with the same two skewed VTCs. To accurately model the failure rate of a custom-designed logic path, we would plot VTCs of all gates and trace the signal propagation through the path. Exact modeling is not possible for standard cell design where the target circuit is unknown. Therefore, although the butterfly plot does not reflect the exact mismatch conditions in a circuit, it does provide a guideline for sizing standard cells consistently to account for local variation.

2.1.2 *Failure Rate From Insufficient Output Swing*

We now define logic failure as having no inscribed square in the butterfly plot and measure how the failure rate varies with V_{DD} and device sizing. To consider logic gates with up to three stacked devices, we verify the INV, NAND2, and NOR2 gates against NAND3 and NOR3, which give the most stringent V_{IH} and V_{IL} requirements respectively. Sizing of NAND3 and NOR3 are fixed to provide a starting point for designing the remaining gates.

The failure rate is estimated from a 5k-point Monte Carlo simulation at worst case temperature. V_T of transistors in the gate under test and global (inter-die) process conditions are randomized such that the Monte Carlo runs are analogous to sampling logic gates across multiple dies. Figure 2(a) shows the failure rate versus V_{DD} of an inverter at various widths normalized to minimum size. Simulated values in markers are fitted to an exponential function ae^{bx}, drawn as a solid line. Note that the failure rate decays more quickly when W=1.66 compared to W=1. Furthermore, zero samples failed in the 5-k point run at higher voltages, as indicated by arrows on the graph.



(a) (b)

Figure 2: Failure rate of (a) inverter and (b) static register vs. V_{DD}, plotted for various NMOS and PMOS widths (normalized to minimum size).

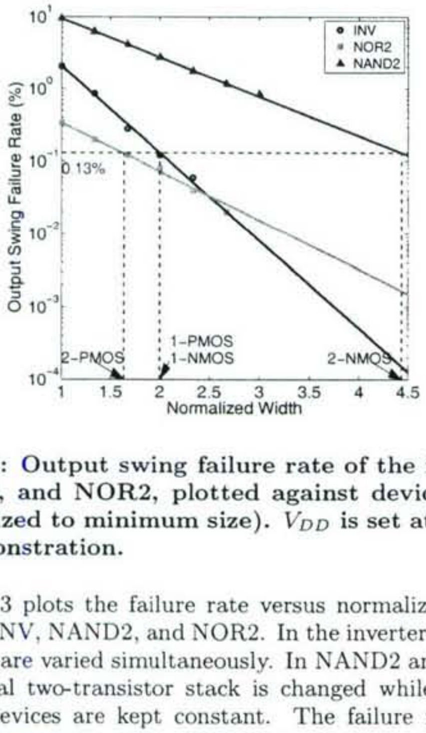


Figure 3: Output swing failure rate of the inverter, NAND2, and NOR2, plotted against device width (normalized to minimum size). V_{DD} is set at 240mV for demonstration.

Figure 3 plots the failure rate versus normalized device width of INV, NAND2, and NOR2. In the inverter, both device sizes are varied simultaneously. In NAND2 and NOR2, the critical two-transistor stack is changed while the two parallel devices are kept constant. The failure rates also decay exponentially with widths. By increasing the device width or V_{DD}, the failure rate can be made to approach 0.

2.2 Noise Margin in Registers

The concept of noise margin is also relevant in sub-threshold register design, where data retention is a particular challenge. Dynamic registers suffer from charge leakage, which worsens in sub-threshold due to slow circuit speeds. Therefore, we consider the static transmission-gate based register. Similar to SRAM cells, the data retention capability of the register is reflected in the hold static noise margin of its cross-coupled inverters. Figure 4 shows the equivalent circuit for measuring the register SNM, accounting for the voltage drop across T2 and the worst case leakage across T1. This circuit is used in a Monte Carlo simulation while varying the V_T of each transistor and inter-die process conditions. Figure 2(b) plots the resulting failure rate in the cross-coupled inverters. Similar to the case of logic gates, the failure rate decreases exponentially to zero when either width or V_{DD} is increased.

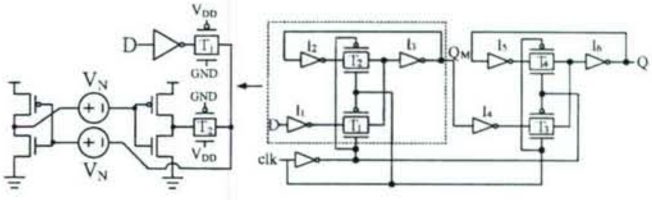


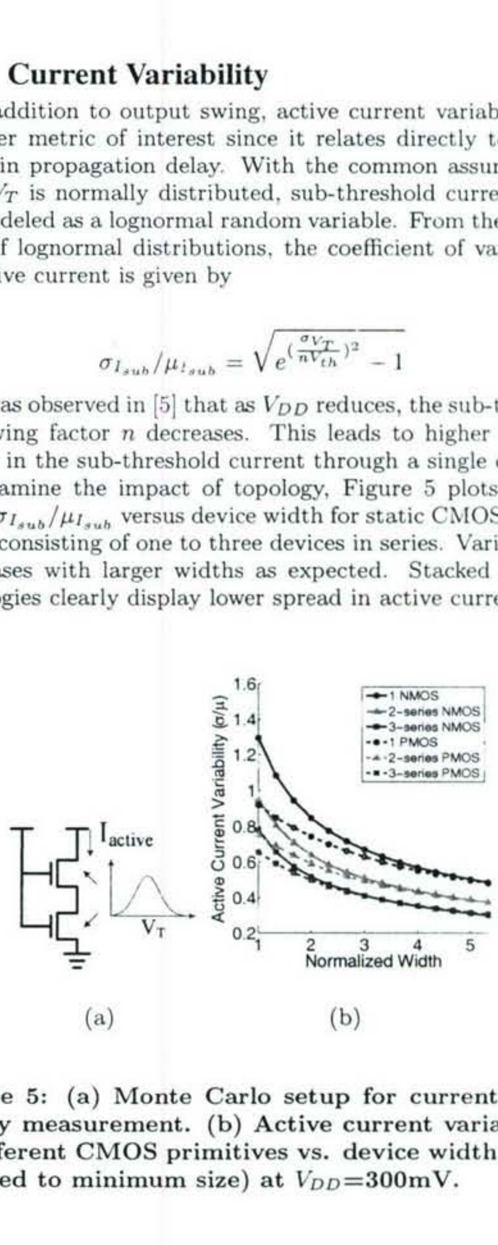
Figure 4: Static register schematic and equivalent circuit for measuring SNM.

2.3 Current Variability

In addition to output swing, active current variability is another metric of interest since it relates directly to variation in propagation delay. With the common assumption that V_T is normally distributed, sub-threshold current can be modeled as a lognormal random variable. From the property of lognormal distributions, the coefficient of variation of active current is given by

$$\sigma_{I_{sub}}/\mu_{I_{sub}} = \sqrt{e^{(\frac{\sigma_{V_T}}{nV_{th}})^2} - 1} \tag{3}$$

It was observed in [5] that as V_{DD} reduces, the sub-threshold swing factor n decreases. This leads to higher uncertainty in the sub-threshold current through a single device. To examine the impact of topology, Figure 5 plots simulated $\sigma_{I_{sub}}/\mu_{I_{sub}}$ versus device width for static CMOS primitives consisting of one to three devices in series. Variability decreases with larger widths as expected. Stacked device topologies clearly display lower spread in active currents.



(a) (b)

Figure 5: (a) Monte Carlo setup for current variability measurement. (b) Active current variability of different CMOS primitives vs. device width (normalized to minimum size) at V_{DD}=300mV.

2.4 Constant Yield Device Sizing

We now address the issue of device sizing for single and stacked device topologies, given the metrics of output swing and current variability. In above-threshold design, series devices are sized to give equivalent resistance as the inverter. However, in sub-threshold design when the objective is to minimize energy, device sizes should be kept as small as possible while satisfying variability constraints.

Compared to a single device, stacked devices display lower current spread but higher uncertainty in output levels, which may lead to functional errors. Reducing the error rate clearly takes precedence, so output swing rather than current variability should be considered first in sizing decisions.

The output swing failure rate versus width plot of Figure 3 illustrates a sizing methodology for single and stacked devices. Suppose we constrain all topologies to have the same failure rate, or interchangeably, a constant yield. We obtain the required device sizes by drawing a horizontal line at the desired failure rate, then finding where this line intersects the failure curve and the corresponding x-axis value. In Figure 3, a target failure rate of 0.13% requires a single and 2-stack NMOS to be sized at 2 and 4.43 times minimum width respectively. 1-PMOS is sized the same as 1-NMOS as both devices are varied together in simulation. The 2-stack sizing here can be used for any static CMOS gate with two series NMOS, since it was derived from NAND2 where two leaking parallel PMOS give the worst case V_{OL}.

Because the failure rate reduces at higher V_{DD}, the required size for a given yield constraint also decreases. The resulting energy trade-off will be analyzed in Section 3.1. Table 1 lists device widths for a constant failure rate of 0.13% while V_{DD} is varied at 20mV intervals. 0.13% represents the 3σ tail of a normal distribution and is chosen for demonstration. It should be noted that such a target allows sizing logic gates consistently, but does not relate in a straightforward way to the failure rate of a circuit built from these gates. As mentioned previously, this value is a pessimistic estimate because it assumes that every second gate in the circuit is NAND3 or NOR3. Furthermore, failing logic gates tend to cluster on die at process corners.

Table 1: Required widths (normalized to minimum size) vs. V_{DD} for constant failure rate=0.13%

$V_{DD}(V)$	0.24	0.26	0.28	0.30	0.32	0.34
1-NMOS	2	1.67	1.33	1	1	1
2-NMOS	4.43	2.93	2.3	2.27	1.3	1
1-PMOS	2	1.67	1.33	1	1	1
2-PMOS	1.63	1	1	1	1	1

3. MINIMUM ENERGY OPERATION

The total energy per operation consumed by an arbitrary circuit is modeled in [2] as

$$E_T = E_{DYN} + E_L = C_{eff}V_{DD}^2 + W_{eff}I_{leak}V_{DD}t_dL_{DP} \quad (4)$$

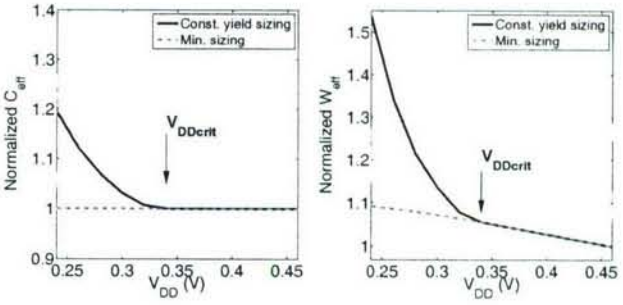
E_{DYN} and E_L model the dynamic switching and leakage energy per cycle respectively. C_{eff} and W_{eff} denote the average total switched capacitance and normalized width contributing to leakage current. t_d and I_{leak} represent the delay and leakage current of a characteristic inverter, while L_{DP} is the logic depth in terms of the inverter delay. As V_{DD} decreases, E_{DYN} is lowered quadratically. The leakage current reduces because of DIBL, but t_d goes up exponentially at sub-threshold voltages and causes a similar increase in leakage energy. The two opposing trends give rise to an optimal supply voltage V_{DDopt} at which total energy is minimized, assuming the circuit is functional.

Section 2 has shown that functionality is no longer guaranteed at low supply voltages when V_T variation is significant. Reducing the probability of logic failure requires either upsizing devices or increasing V_{DD}, which must be considered when finding V_{DDopt}. This can be accounted for within the framework of [2] by treating C_{eff} and W_{eff} as a function of V_{DD}. The resulting energy versus V_{DD} characteristic of an inverter chain and 32-bit Kogge-Stone adder are simulated in a 65nm process and presented as examples.

3.1 Minimum Energy Point with Yield Constraint

Figure 6 plots C_{eff} and W_{eff} versus V_{DD} for the Kogge-Stone adder under two sizing schemes. The solid line plots energy of designs satisfying an upper bound on the output swing failure rate, derived from constant yield sizing of Table 1. The dashed line indicates an adder with only minimum size devices. Note that W_{eff} is obtained by normalizing the adder leakage current to that of a characteristic inverter [2]. DIBL affects leakage through the two circuits differently as V_{DD} decreases, causing a slight increase in W_{eff} in this case. V_{DDcrit} denotes the critical operating voltage at which minimum size devices can be used to satisfy the yield constraint. When $V_{DD} \geq V_{DDcrit}$, the circuit under both schemes are identical.

It should be noted that once the yield constraint is set, V_{DDcrit} can be found immediately from Table 1 and the topology of a given circuit. For example, a circuit without stacked devices does not require upsizing when $V_{DD} \geq V_{DDcrit} = 300$mV. In contrast, a circuit with stacks of two NMOS has $V_{DDcrit} = 340$mV.



(a)

(b)

Figure 6: (a) C_{eff} and (b) W_{eff} for adder with constant yield (CY) and minimum sizing (MS).

The switching, leakage, and total energy of the inverter chain and adder are then calculated according to Equation 4. Figure 7(a) plots the energy versus V_{DD} characteristic of the inverter chain at nominal process and temperature. Total energy in both constant yield and minimum sized chains are dominated by the dynamic component. Therefore, the optimum supply voltage of the minimum size chain (dashed line) is the lowest V_{DD} at which yield constraints are met. By definition, this is equal to V_{DDcrit}. In the constant yield sizing scheme (solid line), reducing the supply below V_{DDcrit} necessitates an increase in device widths. The resulting rise in C_{eff} dominates total energy. In this situation, there is no benefit from upsizing in order to operate at lower V_{DD}. The optimum operating point is with minimum sizing at the lowest V_{DD} permitted by the failure rate constraint.

When the minimum size circuit does have a local minimum in its energy characteristic, three scenarios exist depending on the relationship between V_{DDcrit} and the optimum V_{DD} of the constant yield ($V_{DDopt-CY}$) and minimum sizing ($V_{DDopt-MS}$) schemes.

Case 1) $V_{DDopt-MS} > V_{DDcrit}$: No upsizing is required to operate at the minimum energy point, therefore a minimum sized circuit at $V_{DDopt-MS}$ yields optimum energy.

Case 2) $V_{DDopt-MS} < V_{DDopt-CY} < V_{DDcrit}$: A minimum size circuit cannot operate at $V_{DDopt-MS}$ without violating failure rate constraints. A circuit suitably upsized to operate at $V_{DDopt-CY}$ yields optimum energy while satisfying yield requirements.

Case 3) $V_{DDopt-MS} < V_{DDopt-CY} = V_{DDcrit}$: At V_{DDcrit}, the circuit under both sizing schemes are identical. Therefore a minimum size circuit operating at V_{DDcrit} provides minimum energy.

An example of case 2 is seen in Figure 7(b) for a synthesized 32-bit Kogge-Stone adder with interconnect parasitics extracted from layout. Ignoring failure rate constraints, the minimum size adder (dashed line) has an optimum supply voltage of $V_{DDopt-MS} = 280$mV. When we account for failure rate constraints, the effect of constant yield sizing (solid line) is to add energy overhead when $V_{DD} < V_{DDcrit}$. This shifts the local minimum to the right, hence $V_{DDopt-CY} > V_{DDopt-MS}$. Here $V_{DDopt-CY}$ is also $< V_{DDcrit}$, therefore the adder with constant yield sizing at $V_{DDopt-CY} = 300$mV consumes 10.1% less energy than a minimum size adder at $V_{DDcrit} = 340$mV. In this example, constant yield sizing results in a small reduction in energy due to the shallow minimum of the energy versus V_{DD} curve.

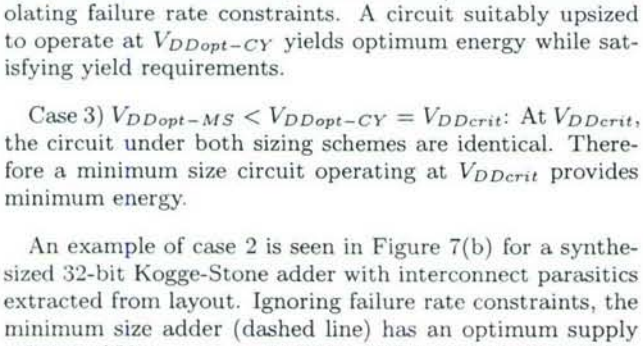


Figure 7: Energy vs. V_{DD} of (a) 11-stage inverter chain and (b) 32-bit adder. Solid and dashed lines indicate CY and MS sizing respectively.

4. PERFORMANCE VARIABILITY

4.1 Delay Variability Modeling

Circuits in sub-threshold display significantly higher delay variability than in above-threshold, therefore proper modeling is essential for timing verification. This section presents a methodology to efficiently model the delay distribution of a chain of logic gates. Using this model, we characterize the delay variability of the Kogge-Stone adders of Section 3.1.

From [2], the delay of a sub-threshold logic gate can be modeled as

$$t_d = \frac{KC_gV_{DD}}{I_oe^{\frac{V_{GS}-V_T}{nV_{th}}}} \tag{5}$$

where K is a delay fitting parameter, C_g is the output capacitance, and the denominator models the gate active current. Both the active current and t_d are lognormally distributed with the same σ parameter. Therefore, delay variability is also given by Equation 3. It depends on σ_{V_T}, which decreases as $(WL)^{-\frac{1}{2}}$, and the sub-threshold swing n, which decreases with V_{DS}. To the first order, σ/μ does not depend on input slew or load capacitance.

The critical path delay in sub-threshold is a sum of lognormal random variables (RVs), typically approximated as another lognormal RV. Authors of [5] derived an expression for the propagation delay of a chain of identical inverters using the Wilkinson approximation. Here we employ the Schwartz-Yeh method [14] to model the sum of non-identically distributed lognormal RVs. The delay of an arbitrary critical path can then be obtained by summing the pre-characterized distributions of each logic gate in the path.

The Schwartz-Yeh method is an iterative algorithm for calculating the sum of lognormal RVs, but requiring much less computation time than Monte Carlo simulation. The modeling methodology using this algorithm is described as follows:

1) Characterize mean delay and standard deviation (μ_{gate}, σ_{gate}) of each logic gate in a cell library, under one input slew and output load condition.

2) Simulate the (N-stage) critical path of interest at nominal process corner and without V_T variation. The delay of the j^{th} stage in the critical path gives μ_{j-path}, for j=1 to N.

3) For each gate j in the critical path, let $\sigma_{j-path} = \sigma_{j-gate} \times \mu_{j-path}/\mu_{j-gate}$, where σ_{j-gate} and μ_{j-gate} are characterized in 1). Since the delay variability σ_j/μ_j is approximately constant across input slew and load conditions, this scales the pre-characterized standard deviation of each gate to the input slew and load conditions in the actual critical path.

4) μ_{j-path} and σ_{j-path} characterize the distribution of each stage, and are input to the Schwartz-Yeh algorithm to generate the delay distribution of the entire critical path.

The above methodology is applied to a three-stage chain consisting of INV-NAND-NOR and to the critical path of a 32-bit Kogge Stone adder at 300mV. Table 2 compares statistical model results with a 1-k point Monte Carlo simulation randomizing V_T of all transistors. The model estimates the mean and standard deviation of the path delay to within a few percent of the Monte Carlo results. This shows that keeping σ/μ constant provides a good approximation.

Table 2: Delay distribution parameters from statistical model and Monte Carlo simulation at 300mV. Values are normalized to FO4 delay.

	Model	Monte Carlo	% Difference
INV-NAND-NOR Chain			
μ	4.957	4.692	5.65%
σ	1.561	1.493	4.51%
Kogge-Stone Critical Path			
μ	36.52	37.13	1.65%
σ	7.038	7.262	3.09%

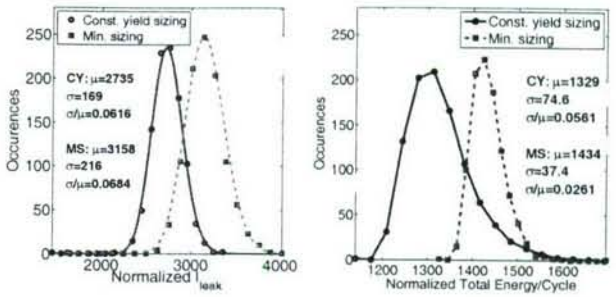
This method is used to characterize the delay distribution of 1) 32-bit adder with constant yield sizing at $V_{DDopt-CY}$ = 300mV, and 2) adder with minimum size devices at V_{DDcrit} = 340mV. Table 3 shows that the first adder exhibits larger mean and 3σ delay, since $V_{DDopt-CY} < V_{DDcrit}$. However, the delay variability of both adders are comparable, indicating that upsized devices in the first adder offset increased variability from operating at a lower supply voltage.

Table 3: Delay distribution comparison of two adders from Section 3.1. Values are normalized to FO4 inverter delay at $V_{DDopt-CY}$.

	Const. Yield Sizing	Min. Sizing
μ	90.88	44.92
σ	17.46	8.857
$\mu + 3\sigma$	143.3	71.49
σ/μ	0.1921	0.1972

4.2 Energy Variability

From a 1k-point Monte Carlo simulation, we characterize the energy distribution of the adder with constant yield sizing at $V_{DDopt-CY}$ and the other with minimum size devices at V_{DDcrit}. As suggested in [5], the switched capacitance is verified to vary negligibly with V_T mismatch and is treated as deterministic. Figure 8(a) shows that even though the former adder employs larger devices, it displays lower mean leakage current due to DIBL, and lower variability as an additional benefit. The first adder exhibits lower mean total energy but higher variability in Figure 8(b). The latter effect results from the delay term in leakage energy having larger mean and standard deviation at 300mV compared to 340mV. Note that the leakage component is a product of two dependent lognormal RVs, so E_T is not strictly lognormally distributed.



(a) (b)

Figure 8: (a) Leakage current and (b) total energy for two adders of Section 3.1, normalized to those of characteristic inverter at $V_{DDopt-CY}$.

5. CONCLUSION

In this paper, we have examined the effect of variation and sizing on single and stacked device topologies in subthreshold circuits. Compared to a single device, stacked devices exhibit lower current variability but a higher probability of logic failure from insufficient output swing. We introduced the use of butterfly plots to verify logic gates as well as registers against process variation, and showed that upsizing is necessary to mitigate degraded output levels. The need for upsizing to meet a given yield constraint imposes an energy overhead and impacts the optimum sizing and supply voltage at which energy is minimized. We presented a methodology to model delay variation in an arbitrary critical path using the delay distribution of each stage. Finally, we compared the delay and energy variability of the proposed sizing scheme with a minimum size circuit, and showed that energy reduction is possible without compromising yield or performance variability.

6. ACKNOWLEDGEMENTS

This research is supported by DARPA and Texas Instruments. The authors are grateful to N. Verma, B. Ginsburg, and D. Finchelstein for helpful discussions.

7. REFERENCES

[1] A. Wang and A. Chandrakasan, "A 180mV FFT Processor Using Sub-threshold Circuit Techniques," in *ISSCC*, 2004, pp. 292–293.

[2] B. H. Calhoun and A. Chandrakasan, "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits," in *ISLPED*, 2004, pp. 90–95.

[3] B. Zhai, *et al.*, "Theoretical and Practical Limits of Dynamic Voltage Scaling," in *DAC*, 2004, pp. 868–873.

[4] B. H. Calhoun, *et al.*, "Device Sizing for Minimum Energy Operation in Subthreshold Circuits," in *CICC*, Oct. 2004, pp. 95–98.

[5] B. Zhai, *et al.*, "Analysis and Mitigation of Variability in Subthreshold Design," in *ISLPED*, 2005, pp. 20–25.

[6] Y. Cao and L. T. Clark, "Mapping Statistical Process Variations Toward Circuit Performance Variability: An Analytical Modeling Approach," in *DAC*, June 2005, pp. 658–663.

[7] S. H. Choi, *et al.*, "Novel Sizing Algorithm for Yield Improvement under Process Variation in Nanometer Technology," in *DAC*, 2004, pp. 454–459.

[8] J. Chen, *et al.*, "Robust Design of High Fan-In/Out Subthreshold Circuits," in *IEEE Int. Conf. on Computer Design (ICCD)*, Oct. 2005, pp. 405–410.

[9] ——, "Maximum-Ultra-Low Voltage Circuit Design in the Presence of Variations," in *IEEE Circuits and Devices Magazine*, Jan.-Feb. 2006, pp. 12–20.

[10] N. Jayakumar and S. P. Khatri, "A Variation-tolerant Sub-threshold Design Approach," in *DAC*, 2005, pp. 716–719.

[11] V. De, *et al.*, "Techniques for Leakage Power Reduction," in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, *et al.*, Eds. IEEE Press, 2001, ch. 3, pp. 46–62.

[12] E. Seevinck, *et al.*, "Static Noise Margin Analysis of MOS SRAM Cells," *IEEE J. Solid-State Circuits*, vol. SC-22, no. 5, pp. 748–754, Oct. 1987.

[13] J. Lohstroh, *et al.*, "Worst-case Static Noise Margin Criteria for Logic Circuits and Their Mathematical Equivalence," *IEEE J. Solid-State Circuits*, vol. SC-18, no. 6, pp. 803–807, June 1983.

[14] S. Schwartz and Y. Yeh, "On the Distribution Function and Moments of Power Sums with Log-Normal Components," *Bell Sys. Tech. Journal*, vol. 61, no. 7, pp. 1441–1462, Sept. 1982.

Sub-Threshold Design: The Challenges of Minimizing Circuit Energy

B. H. Calhoun[1], A. Wang[2], N. Verma[3], and A. Chandrakasan[3]
[1]University of Virginia, [2]Texas Instruments, [3]Massachussetts Institute of Technology
bcalhoun@virginia.edu, aliwang@ti.com, {nverma,anantha}@mtl.mit.edu

ABSTRACT

In this paper, we identify the key challenges that oppose sub-threshold circuit design and describe fabricated chips that verify techniques for overcoming the challenges.

Categories and Subject Descriptors

B.7.1 [**ICs**]: Types and Design Styles

General Terms

Performance, Design, Reliability

Keywords

Sub-threshold digital circuits, low voltage memory, dynamic voltage scaling, process variations, sub-threshold logic

1. INTRODUCTION

Sub-threshold operation for digital circuits first was shown as the means to minimizing CMOS V_{DD} in 1972 [1]. Analog sub-threshold circuits subsequently received a lot of attention for low power applications (e.g. [2][3]). Interest in digital sub-threshold was revived in the late 1990s [4], and a multiplier was demonstrated operating in sub-V_T at 0.475V that used body bias to balance p/n currents [5]. A sub-V_T ring oscillator also employed body biasing and functioned at 80mV [6].

The primary motivation for using sub-V_T circuits is to reduce energy. Analysis of energy contours in [7] demonstrated that minimum energy operation occurs in the sub-threshold region. Once $V_{DD}<V_T$, delay increases exponentially with additional voltage scaling. Leakage current integrates over the longer delay until leakage energy per operation exceeds the active energy and causes the minimum point. Models capture this effect and illustrate the impact of various parameters in [8][9].

The potential for minimizing energy at the cost of speed degradation defines the set of applications for which sub-threshold circuits are well-suited. First, energy-constrained applications such as wireless sensor nodes, RFID tags, or implants are dominated by the need to minimize energy consumption. Speed is a secondary consideration for this class of applications, so sub-V_T circuits offer a good solution. Secondly, many burst-mode applications require high performance for brief time periods between extended sections of low performance operation. Sub-threshold circuits can minimize energy for computations executed during the low performance slots. Finally, the parallelism inherent in many signal processing and communications circuits can be exploited to scale voltages into sub-V_T, providing a low energy solution for throughput-centric applications (e.g. [10]).

This paper describes the key challenges that confront sub-threshold circuit designers and presents chips that overcome the challenges.

2. Sub-Threshold Logic: FFT Processor

Static CMOS gates continue to function in sub-V_T, but some challenges make logic design more difficult. First, CMOS processes are designed with strong-inversion operation in mind, so the ratio of drive current in sub-V_T is frequently imbalanced relative to the case where pMOS and nMOS are symmetrical. The shaded region in Figure 1 shows the operational range for a ring oscillator in 0.18μm CMOS at the worst-case corners. V_{DD} is minimized when the p/n sizing ratio is 12, which indicates that the process is imbalanced such that p/n current is 1/12 relative to the symmetric case. This unfriendly sort of technology imbalance can aggravate process variations and even require different circuit designs for different imbalance scenarios. In addition, the low V_{DD} results in a reduced I_{on}/I_{off} ratio that can reduce robustness, especially for circuits with parallel leakage paths [11].

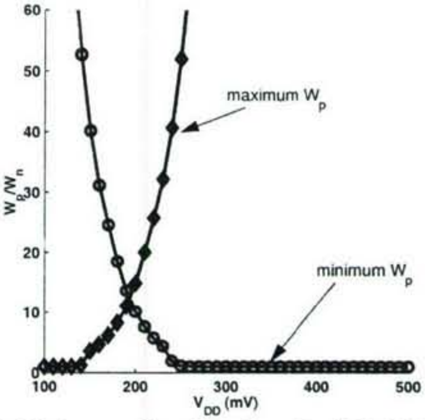


Figure 1: Minimum achievable voltage for 10%-90% output swing for 0.18μm ring oscillator at worst case process corners (simulation).

A 0.18μm CMOS FFT processor uses circuits that account for these challenges: static CMOS logic is used for robustness, gates with parallel leakage paths are redesigned, large stacks are avoided to improve I_{on}/I_{off}, and a register-file memory uses logic-based structures. The chip is fully functional for 128, 256, 512, and 1024 FFT lengths (8-bit and 16-bit precision) at V_{DD} from 180mV to 900mV [11]. Figure 2 shows the measured energy consumption for 8-bit and 16-bit processing as a function of voltage. 8-bit processing has a lower activity factor and thus has lower switching energy. However, because the leakage energy is the same for both 8-bit and 16-bit processing, the minimum

ISLPED'06, October 4–6, 2006, Tegernsee, Germany.
Copyright 2006 ACM 1-59593-462-6/06/0010...$5.00.

energy point increases to 400mV from 350mV. At the 16-bit optimum, the chip runs at 10 kHz and consumes 155nJ/FFT, which is 350X more energy efficient than a typical low-power microprocessor and 8X more energy efficient than a standard ASIC implementation [11].

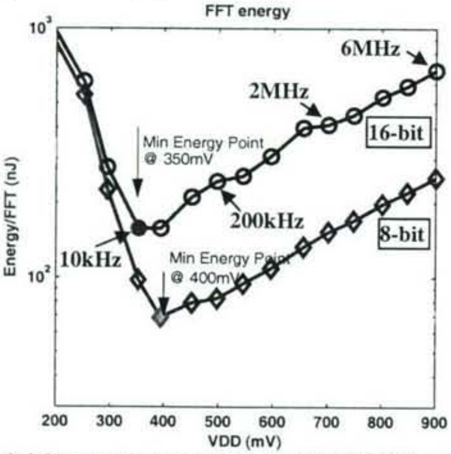


Figure 2: Measured energy per 8- and 16-bit FFT vs. V_{DD}.

3. Scaling Performance: Ultra-DVS

Burst mode applications cannot exclusively utilize sub-threshold operation because they require periodic high speed functionality. Traditional dynamic voltage scaling (DVS) could be extended to include sub-threshold operation, but the overhead of providing the necessary voltages can be large. Adjustable DC-DC converters tend to have limited efficiency over broad voltage ranges, and they take 100s of micro-seconds to switch. An alternative implementation method called local voltage dithering (LVD) offers a reduced overhead means for implementing ultra-DVS (UDVS) down to the sub-threshold region. LVD uses power switches to select from among two or more V_{DD} supplies at the local block level [12]. Figure 3 shows an example system that has 3 V_{DD}s. As the required rate (normalized frequency) for processing incoming data changes, each block spends a different fraction of its operating time at different voltage levels. The averaging effect of this dithering produces an energy consumption profile that nears the optimal (e.g. infinite voltage levels) profile.

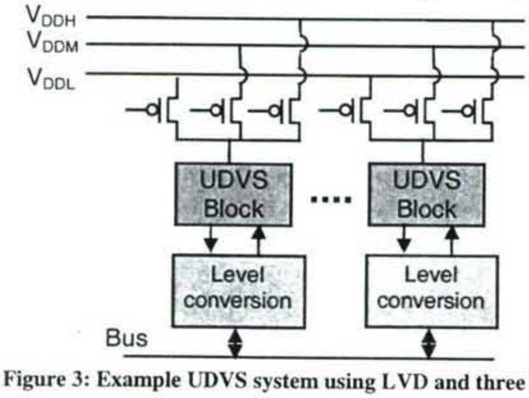


Figure 3: Example UDVS system using LVD and three V_{DD}s.

A 90nm CMOS test chip uses LVD to implement UDVS for 32-bit Kogge-Stone adders [12]. Measurements from the chip show that high rate (e.g. >0.1) dithering can occur in 1 cycle due to the local granularity of the headers. Figure 4 shows an example energy profile for a UDVS system using energy measurements from the test chip. For high rates, the blocks dither between the top two supplies (1.1V and 0.8V in the figure) to achieve near-optimal energy consumption. When performance requirements relax for low rate operation, the blocks can hop to the V_{DD} that gives minimum energy operation (330mV for the 90nm adder block) to achieve 9X savings in energy consumption.

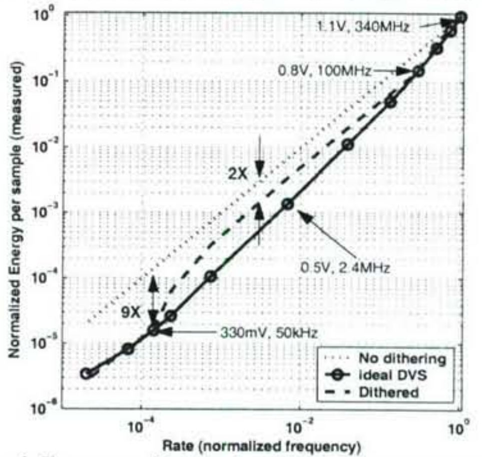


Figure 4: Energy profile based on 90nm chip measurements for example 3-V_{DD} system.

4. Sub-Threshold SRAM

SRAM is an important component of many ICs, and it can contribute a large fraction of the active and leakage power consumption. It is important to have sub-V_T compatible SRAMs for sub-V_T systems. However, the nature of SRAM circuits makes them a melting pot of all of the major sub-V_T challenges.

Random variation fundamentally affects the geometry and threshold voltage of CMOS devices and is increasingly prominent in scaled technologies. The large array nature of SRAM implies that extreme tails of the distributions limit yield. The problem is exacerbated in sub-V_T, where device strength depends exponentially on threshold voltage, and, in the presence of variation, relative strengths cannot be guaranteed by sizing. As a result, the widely used 6T SRAM cell, which relies on ratioed operation and is used to maintain density, fails to operate in sub-V_T. Figure 5a,b show the read/hold and write static nose margins [13] respectively for a typical 6T cell and for the 3σ case. At reduced voltages, read margin is negative and write margin is positive, indicating failure for both operations.

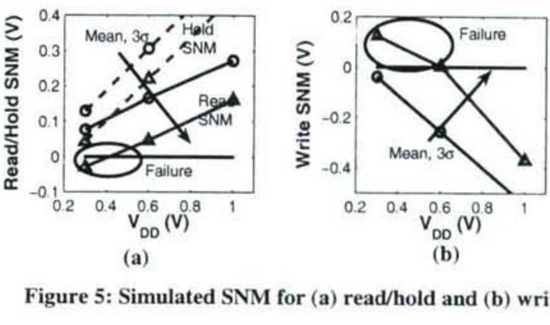


Figure 5: Simulated SNM for (a) read/hold and (b) write.

The increased impact of variation on device strength in sub-V_T also has a limiting effect on SRAM performance and integration. SRAM cell read current, I_{RD}, decreases exponentially in sub-V_T, but the speed is ultimately set by the weakest cell in the array. Figure 6a plots I_{RD} for cells on the weak side of the distribution normalized to the mean (i.e. $I_{RD}/\mu(I_{RD})$). The limiting effect of cell strength variation is amplified in sub-V_T where cells can be over an order of magnitude weaker than the mean.

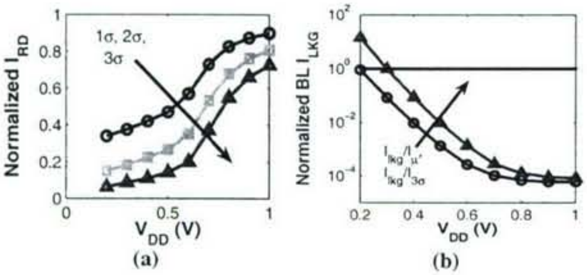


Figure 6: Effect of cell variation on (a) worst case read current and (b) bit-line leakage.

Parallel leakage also limits voltage scaling for SRAM. In conventional 6T SRAM, a stored "1" is read dynamically from a precharged bit-line. However, the reduced I_{on}/I_{off} ratio in sub-V_T is lowered even more due to the unaccessed cells sharing the bit-line, which results in a degraded logic level. Sub-V_T bit-line leakage is less problematic at high voltages where the discharge time of an accessed cell is much faster than that of the aggregate unaccessed cells. However, where variation extends the required discharge time, bit-line leakage severely limits the number of cells that can be integrated onto a column. Figure 6b shows the leakage current of 127 unaccessed cells normalized to the drive current of a single accessed cell weakened by variation. Values greater than unity, which occur in sub-V_T, imply that drive current is indistinguishable from leakage, making reliable read accesses impossible.

Numerous techniques have been reported to mitigate the low-voltage SRAM problems described above. For instance, reduced bit-line precharge voltages and negative word-line bias for unaccessed cells have been used to increase the read SNM. Similarly, increased word-line bias and negative bit-line voltages have been used to improve the write SNM. While these approaches can improve the situation for sub-V_T SRAM, approaches that address the problems more fundamentally provide a better solution for robust operation in sub-threshold.

A 65nm test chip implements a 256kb memory that overcomes the problems and provides functionality in the sub-threshold region to below 400mV [14]. The SRAM uses a 10T bit-cell, shown in Figure 7. M7-M10 form a read buffer that isolates the internal storage nodes, Q and QB, so that a read upset is not possible. This eliminates the read SNM problem of Figure 5a, and stability is instead limited by the hold SNM. Measurements from the test chip show that the cell can hold data correctly below 250mV. Write operations in Figure 5b fail since the access devices in a 6T bit-cell are too weak to over-power the internal cell feedback, which is made worse by process imbalance that makes pMOS sub-threshold current higher than nMOS by an order of magnitude. Robust write in the new 10T cell is performed by weakening the feedback structure by floating VV_{DD}. Finally, bit-line leakage on RBL is minimized by unconditionally raising the voltage of QBB for unaccessed cells. This relies on either the active pull-up current through M9, or the ratio of its leakage current to that of M10's. In either case, M8's V_{GS} becomes negative, resulting in vanishingly small sub-threshold leakage current to the bit-line. This structure allows 256 bit-cells to be integrated per column.

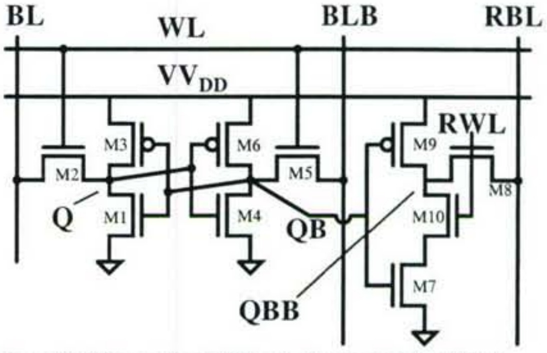


Figure 7: Schematic of 10T sub-threshold bit-cell [14].

5. Conclusions

Numerous problems increase the challenge of designing robust sub-threshold circuits. Some time-testing design practices, such as ratioed write in SRAM, become unreliable due to the exponential dependence of sub-threshold drive current on parameters with large process variations. We have presented an overview of the types of circuits and architectures that overcome these problems and produce working designs. Functional implementations of a sub-threshold FFT processor [11], an energy-scalable UDVS test chip [12], and a sub-threshold SRAM [14] attest that robust sub-threshold systems can practically offer minimum energy operation.

6. ACKNOWLEDGEMENTS

We acknowledge DARPA and Texas Instruments for funding.

7. REFERENCES

[1] Swanson and Meindl, *JSSC*, 1972.
[2] Vittoz and Fellrath, *JSSC*, 1977.
[3] Mead, Addison-Wesley, 1989.
[4] Soeleman and Roy, *ISLPED*, 1999.
[5] Paul, Soeleman, and Roy, *ESSCIRC*, 2001.
[6] Deen, Kazemeini, and Naseh, *ICCDCS*, 2002.
[7] Wang, Chandrakasan, and Kosonocky, *SVLSI*, 2002.
[8] Zhai, Blaauw, Sylvester, and Flautner, *DAC*, 2004.
[9] Calhoun and Chandrakasan, *ISLPED*, 2004.
[10] Sze, Blazquez, Bhardwaj, and Chandrakasan, *ICASSP*, 2006.
[11] Wang and Chandrakasan, *ISSCC*, 2004.
[12] Calhoun and Chandrakasan, *ISSCC*, 2005.
[13] Seevinck, List, and Lohstroh, *JSSC*, 1987.
[14] Calhoun and Chandrakasan, *ISSCC*, 2006.

3.2 Minimum Energy Tracking Loop with Embedded DC-DC Converter Delivering Voltages down to 250mV in 65nm CMOS

Yogesh K. Ramadass, Anantha P. Chandrakasan

Massachusetts Institute of Technology, Cambridge, MA

Minimizing the energy consumption of battery powered systems is a key focus in integrated circuit design. Switching energy of digital circuits reduces quadratically as V_{DD} is decreased below V_T (i.e. sub-threshold operation), while the leakage energy increases exponentially. These opposing trends result in a minimum energy point (MEP), defined as the operating voltage at which the total energy consumed per operation (E_{op}) is minimized [1]. Operating circuits at their MEP [1, 2] has been proposed as a solution for energy critical applications and the analytical solution of the MEP has been derived in [3]. The MEP can vary widely for a given circuit depending on its workload and environmental conditions (e.g., temperature). By tracking the MEP as it varies, energy savings of 50 - 100% are demonstrated and even greater savings can be achieved in circuits dominated by leakage. In this paper, a 65nm CMOS circuit that can dynamically track the MEP of a digital circuit with varying operating conditions is presented. Embedded within the tracking loop is an ultra-low-power switching DC-DC converter that can efficiently deliver supply voltages down to 250mV, enabling minimum energy operation.

Figure 3.2.1 shows the architecture of the MEP tracking loop, which adjusts the output V_{DD}, to the minimum energy operating voltage of the digital circuit (FIR filter). An energy sensor circuit, together with an energy minimization algorithm, is used to set the reference voltage of the DC-DC converter. The DC-DC converter maintains V_{DD} close to the reference voltage. The key element in the loop is the energy sensor circuit which computes the E_{op} of the digital circuit at a given reference voltage. The DC-DC converter is disabled during energy sensing. Assuming that the voltage across the storage capacitor C_{load} falls from the reference voltage V_1 to V_2 in the course of N operations of the digital circuit, E_{op} at the voltage V_1 is equal to $C_{load} \times (V_1^2 - V_2^2)/2N$. To measure E_{op} accurately, V_2 should be close in value (within 20 - 30mV) to V_1. Methods to measure E_{op}, by digitizing V_1 and V_2 using conventional ADC's, or by sensing the inductor current, dissipate a significant amount of overhead power. Our proposed energy efficient approach to obtain E_{op} is to observe that, by design, V_1 is very close to V_2. Thus, $V_1^2 - V_2^2$ can be simplified to $2V_1 \times (V_1 - V_2)$ within an acceptable error. Since, the digital representation of V_1, which is the reference voltage to the DC-DC converter, is already known, only the digital value for $V_1 - V_2$ is required to estimate E_{op}.

Figure 3.2.2 shows the voltage difference measuring circuitry. Before starting an N operation energy sense cycle, the voltage V_1 across C_{load} is sampled on C_1 and the DC-DC converter is disabled. The digital circuit runs for N operations using the energy stored in C_{load}, and the voltage across C_{load} droops to some value V_2 (< V_1), which is then sampled across C_2. Subsequently, the DC-DC converter is enabled and normal operation of the digital circuit continues. At this point, a current sink (M_1, M_2) connected across C_1 turns ON and a fixed frequency clock drives a counter. The fixed frequency clock, together with the constant current sink that drains C_1, quantizes voltage into time steps, as in an integrating ADC. The number of fixed frequency clock cycles required for C_1 to droop down to V_2 is directly proportional to $V_1 - V_2$. Once the value of $V_1 - V_2$ is obtained digitally, it is multiplied with V_1 to get an estimate of E_{op}.

The digital representation of E_{op} is then used by a slope descent algorithm to arrive at the MEP. Based on the value of E_{op} obtained, the algorithm suitably changes the reference voltage to the DC-DC converter. Once the converter settles at this new voltage, the energy sensing operation is performed again and the cycle repeats until the minimum is achieved. At this point the loop shuts down. Figure 3.2.3 shows measured waveform of the tracking loop in operation. The MEP tracking loop can be enabled by a system controller as needed depending on the application.

Figure 3.2.4 shows the DC-DC converter embedded within the minimum energy tracking loop. The converter is a synchronous rectifier buck converter [4] with off-chip filter elements. It is designed to deliver load voltages from V_{DD} = 250mV to as high as V_{DD} = 700mV at ultra-low load power levels from 1μW to 100μW. This precludes the usage of high gain amplifiers for zero voltage and current switching. The converter implemented uses an open loop control for zero current switching. Depending on the load voltage being delivered, an appropriate delay is multiplexed in, turning the NMOS off when the inductor current approaches zero (see Fig. 3.2.4). A Pulse Frequency Modulation (PFM) control scheme is used to improve efficiency as the load power levels are low. The clock for the reference voltage comparator is derived from the critical path replica ring oscillator which feeds the digital circuit. This allows the comparator clock to scale automatically with V_{DD} and hence the load power, eliminating unnecessary comparisons. The simplicity of open-loop PFM mode control helps in decreasing the power consumption of the control circuitry, thereby improving the low load efficiency. The converter efficiency, plotted from measured results in Fig. 3.2.5, is >80% while delivering load powers of 1μW and higher and 86% at 100μW (V_{DD}=0.5V).

Figure 3.2.6 shows how the MEP varies with workload for a 7-tap FIR filter implemented in 65nm CMOS. Workload is changed by varying the number of taps of the FIR filter. The MEP decreases with increasing workload because the ratio of the active energy to total energy per operation increases. It can be deduced from curves 1, 2 in Fig. 3.2.6 that 110% energy is saved by moving V_{DD} to the new MEP value instead of staying at the original MEP value of 320mV. The MEP increases with temperature as the ratio of leakage energy to total energy per operation increases. Energy savings of the order of 50% is achieved as the MEP is tracked when the temperature changes from 0 to 85°C. The energy savings obtained are highly circuit dependent and can be much larger in modern digital IC's, which dissipate a significant portion of power in leakage.

The energy overhead associated with obtaining the MEP is equivalent to the energy consumed by 50 operations at the MEP in the minimum workload scenario (WL1). The proposed minimum energy tracking loop is *non-intrusive*, thereby allowing the load circuit to operate without being shut down. The tracking methodology is independent of the size and type of digital circuit being driven and the topology of the DC-DC converter.

Figure 3.2.7 shows the micrograph of the test chip fabricated in a 65nm CMOS process. The active area of the chip, which includes the digital test circuitry, occupies 0.23mm² with the minimum energy tracking circuitry occupying 0.05mm². The small area and energy overhead of the tracking loop facilitates the use of multiple such loops for each distinct voltage domain in a complex digital system.

Acknowledgements:

This work was funded by DARPA. We thank Texas Instruments for chip fabrication.

References:

[1] A. Wang and A. Chandrakasan, "A 180-mV Subthreshold FFT Processor Using a Minimum Energy Design Methodology," *IEEE J. Solid-State Circuits*, vol. 40, pp. 310-319, Jan., 2005.

[2] B. Zhai, et al., "A 2.6pJ/Inst Subthreshold Sensor Processor for Optimal Energy Efficiency," *IEEE Symposium on VLSI Circuits*, pp. 192-193, June, 2006.

[3] B. H. Calhoun and A. Chandrakasan, "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits," *IEEE ISLPED*, pp. 90-95, Aug., 2004.

[4] J. Xiao, et al., "A 4μA-Quiescent-Current Dual-Mode Buck Converter IC for Cellular Phone Applications," *ISSCC Dig. Tech. Papers*, pp. 280-281, Feb., 2004.

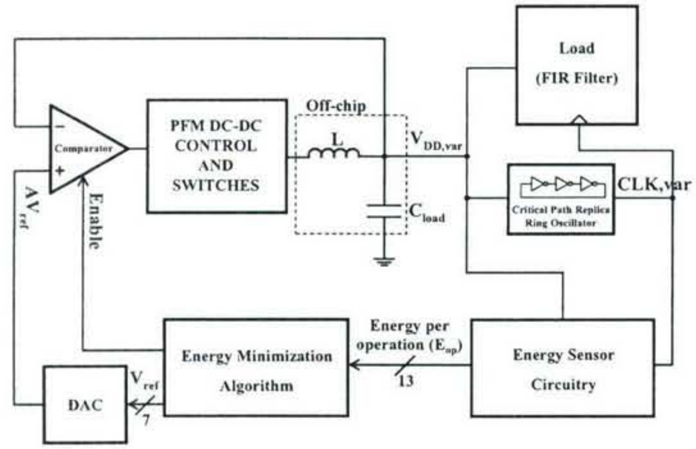

Load
(FIR Filter)
PFM DC-DC CONTROL AND SWITCHES
Comparator
Off-chip
L
$V_{DD,var}$
C_{load}
Critical Path Replica Ring Oscillator
CLK,var
AV_{ref}
Enable
Energy per operation (E_{op})
Energy Minimization Algorithm
Energy Sensor Circuitry
13
DAC
V_{ref}
7

Figure 3.2.1: Block diagram of the minimum energy tracking loop and embedded DC-DC converter.

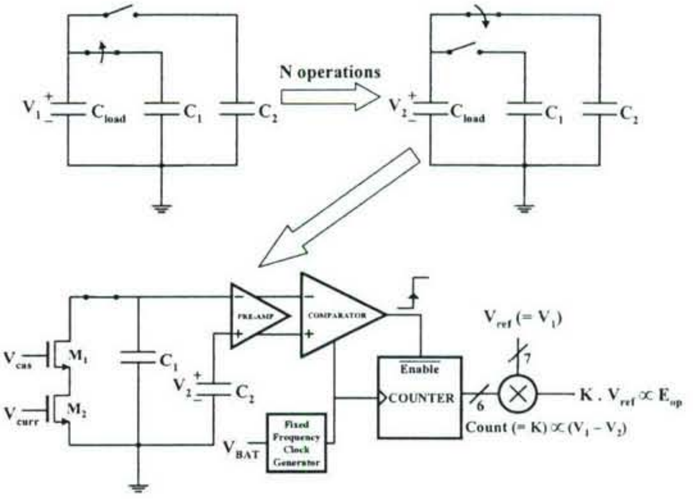


Figure 3.2.2: Circuitry to compute Energy/operation (E_{op}) at a given operating voltage.

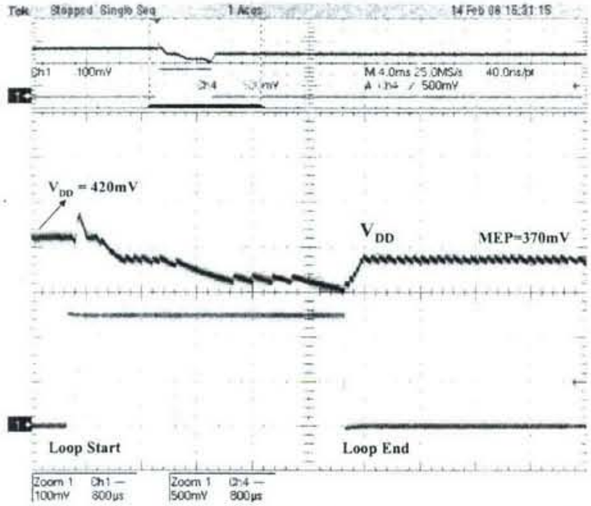


Figure 3.2.3: Measured waveform showing the minimum energy tracking loop in operation. V_{DD} starts at 420mV and is then increased to 470mV. The loop then changes direction and reduces V_{DD} to 370mV and 320mV before settling at the MEP of 370mV.

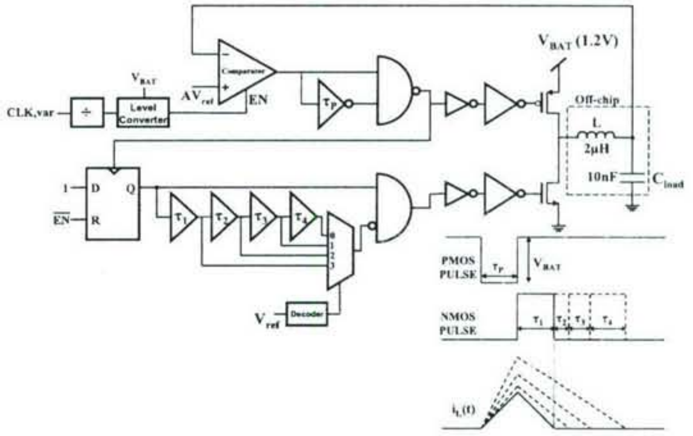


Figure 3.2.4: Pulse Frequency Modulation control of the DC-DC converter showing open-loop NMOS pulse width determining circuitry. The time delays are chosen to turn the NMOS off as the inductor current approaches zero.

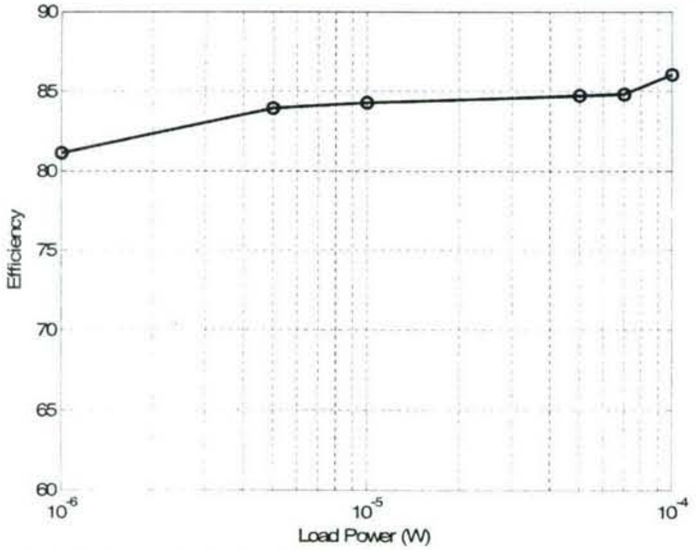


Figure 3.2.5: Measured efficiency plot of the DC-DC converter.

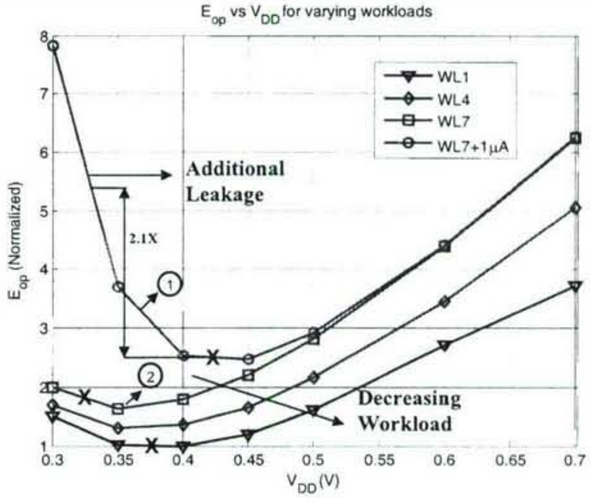


Figure 3.2.6: Measured E_{op} curves with change in workload for a 7-tap FIR filter. Curve 1 has an intentional 1µA leakage current added to the maximum workload scenario (curve 2). 'X' denotes the measured voltage at which the minimum energy loop settles.

Continued on Page 587

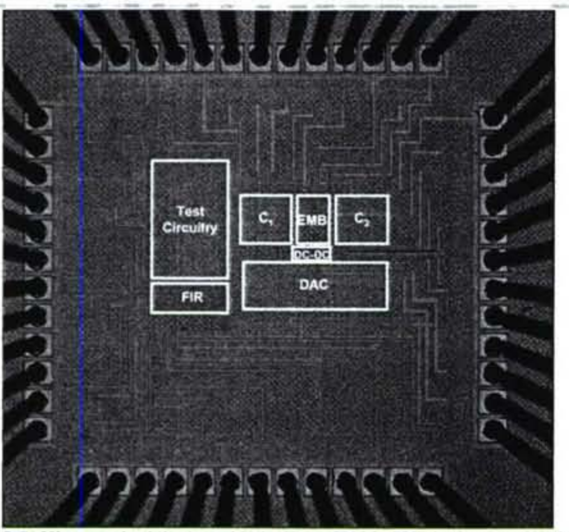


Figure 3.2.7: Micrograph of the test chip in 65nm CMOS. EMB is the energy minimizing block which comprises the energy sensor circuitry and the energy minimization algorithm.

18.4 A 65nm 8T Sub-V$_t$ SRAM Employing Sense-Amplifier Redundancy

Naveen Verma, Anantha P. Chandrakasan

Massachusetts Institute of Technology, Cambridge, MA

The subthreshold regime is a critical biasing space as it enables minimum energy operation for logic circuits [1]. However, practical systems rely heavily on SRAMs, which conventionally limit the minimum V_{DD} to above V_t. SRAMs often dominate the total die area and power, and minimizing their energy requires scaling V_{DD} as low as possible. In this work, a 256kb SRAM in 65nm CMOS is presented that operates in sub-V_t (at 350mV) despite the exponential effect V_t variations have on device strength.

The 6T bit-cell in Fig. 18.4.1 provides a good balance between stability, performance, and density. However, in the presence of variation, it fails to operate in sub-V_t. Figure 18.4.1 shows a Monte Carlo simulation of the SNM [2] for both read and hold cases of a 65nm cell. At 350mV, hold stability is preserved, but read failures are prominent. Write SNM violations (not shown) appear in a similar manner. Functional errors are also caused by severely degraded I_{READ}. Figure 18.4.1 considers the case of 256 cells per column. In sub-V_t, the values stored in the unaccessed cells can result in an aggregate leakage current on the shared bitlines that is greater than the 3σ and 4σ read currents, implying that the data in the accessed cell is indistinguishable from bitline leakage.

To overcome these challenges, the 8T bit-cell shown in Fig. 18.4.2 is developed. Buffered read eliminates the read SNM limitation; peripheral footer circuitry eliminates bitline leakage; peripheral write drivers and storage-cell supply drivers interact to reduce the cell supply voltage during write operations; and sense-amp redundancy provides a favorable trade-off between offset and area. Previous implementations of sub-V_t memories deal with stability, read-current, and bitline leakage by adding devices within the cell or employing hierarchy to limit fan-in/out. For instance, a 10T cell operates at 400mV [3], and a register-file uses multiplexed read to operate at 310mV [4]. In this design, peripheral circuit assists are used to maximize density and reduce the leakage paths to those of a 6T SRAM.

In Fig. 18.4.2, the read buffer is composed of M7-M8. Instead of statically connecting its foot to ground, however, a foot-driver is used in the periphery. As shown in Fig. 18.4.3, the buffer-foots of all cells of the same word are shorted, and their foot-driver is shared. During a read, only the foot of the accessed word is driven low; all others remain at V_{DD}. Accordingly, after RDBL is precharged, the read-buffers of the unaccessed cells have no voltage drop across them, and their access devices have a negative V_{GS}. Consequently, they impose no sub-V_t leakage, and dynamically held data values of "1" on RDBL can be sensed successfully.

The foot-driver is required to sink the read current from all of the accessed cells. Use of a large NMOS to accomplish this is impractical since it would impose a significant area and leakage-power overhead. Instead, the sub-V_t charge-pump circuit shown in Fig. 18.4.3 is used. The voltage boost provided by typical charge-pump implementations suffers from V_t drops, and would be inadequate for this application. Instead, the circuit of Fig. 18.4.3 uses a PMOS (M1) to precharge C_{BOOST}. The charge-pump generates a swing of nearly $2V_{DD}$ at the input of the foot-driver, enhancing its current by over two orders of magnitude while reducing V_t variation dependencies on its devices. This allows the devices of the foot-driver to be near minimum sized so that their leakage-power is insignificant. Further, since the charge-pump drives minimal load, its devices and boost capacitor can be small, consuming negligible power and area.

Write operations fail when the cell pass devices cannot overpower the internal cell feedback. In this design, write (Fig. 18.4.4) is performed by boosting WL by 50mV and, more importantly, reducing VV_{DD} through a supply driver. Simultaneously, new data is written primarily by pulling the desired storage node low through the NMOS pass device. Although, the opposite storage node is only weakly pulled high, its load PMOS provides a current path to VV_{DD}. Accordingly, all cells in the accessed word contribute to driving VV_{DD} high through one of their NMOS pass devices. Relatively large devices are used in the supply driver, and the net variation in the pass devices and write drivers tends to average; hence, sizing accurately allows VV_{DD} to be set to a low intermediate voltage.

The write mechanism, which is essential for sub-V_t operation, requires each word to have a separate VV_{DD}. As shown in Fig. 18.4.5, this implies that columns of different blocks cannot be interleaved in layout, and adjacent columns can no longer share a multiplexed sense-amp. Hence, the number of sense-amps required increases, and each must fit in a column pitch. Nominally, the approach of large-signal read, which is advantageous in high-density, scaled SRAMs [5], is used; nonetheless, the BL voltage levels are degraded, due to gate-leakage and other noise mechanisms, and sense-amp offsets still limit yield. To remedy this, sense-amp redundancy is employed. Erroneous reads occur when the net offset of each sensing network is greater than the input voltage swing. Increasing device sizes reduces local variation, accordingly reducing sense-amp offset. Redundancy, however, allows exclusive selection of the sense-amp that minimizes the achievable offset. Hence, errors now depend on the joint probability that all sense-amps have an offset greater than the input voltage swing. As shown in Fig. 18.4.5, the error probability for a half-sized sense-amp is greater than that for a unit-sized sense-amp; however, Monte Carlo simulation shows that the joint error probability for two half-sized sense-amps is lower than that for a unit-sized sense-amp. Specifically, a factor of five improvement is observed at the input swings of interest (i.e., 50mV). This only applies where the errors due to offset are uncorrelated, so, a pseudo-differential sense-amp structure is employed to cancel the effects of global variation.

Increased redundancy yields further improvement, but the overhead of selecting between redundant sense-amps and storing that selection state also increases. In this design, two sense-amps are used, requiring the minimal support circuitry of two flip-flops and a few logic gates. On start-up a selection routine determines which sense-amp can correctly read both logic "0" and "1", and enables only the corresponding structure.

The SRAM is fabricated in a 65nm CMOS process (Fig. 18.4.7). The 256kb array is arranged into 8, 256 row × 128 column blocks. Full read and write functionality is achieved with a V_{DD} of 350mV (and 50mV boosting of WL drivers). At this voltage, the SRAM operates at 25kHz and consumes 2.83μW during read and 3.96μW during write. As shown in Fig. 18.4.6, data is held to 300mV where the leakage power is 1.92μW. At 325mV fewer than 0.05% read/write errors are observed.

Acknowledgements:
This work is funded by DARPA, and IC fabrication is provided by Texas Instruments. N. Verma is supported by an Intel Ph.D. Fellowship. The authors thank B. Calhoun and M. Bhardwaj for their support and feedback.

References:

[1] A. Wang, A. Chandrakasan, and S. Kosonocky, "Optimal Supply and Threshold Scaling for Sub-Threshold CMOS Circuits," *Proc. IEEE Comp. Society Annual Int. Symp. VLSI*, pp. 5-9, Apr., 2002.

[2] E. Seevinck, F. List, and J. Lohstroh, "Static Noise Margin Analysis of MOS SRAM Cells," *IEEE Journal of Solid-State Circuits*, SC-22, no. 5, pp. 748-754, Oct. 1987.

[3] B. Calhoun and A. Chandrakasan, "A 256kb Sub-Threshold SRAM in 65nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 628-629, Feb., 2006.

[4] J. Chen, L. Clark, and T.-H. Chen, "An Ultra-Low-Power Memory With a Subthreshold Power Supply Voltage," *IEEE J. Solid-State Circuits*, vol. 41, no. 10, pp. 2344-2353, Oct., 2006.

[5] K. Zhang, K Hose, V. De, et al., "The Scaling of Data Sensing Schemes for High-Speed Cache Design in Sub-0.18μm Technologies," *Symp. VLSI Circuits*, pp. 226-227, Jun., 2000.

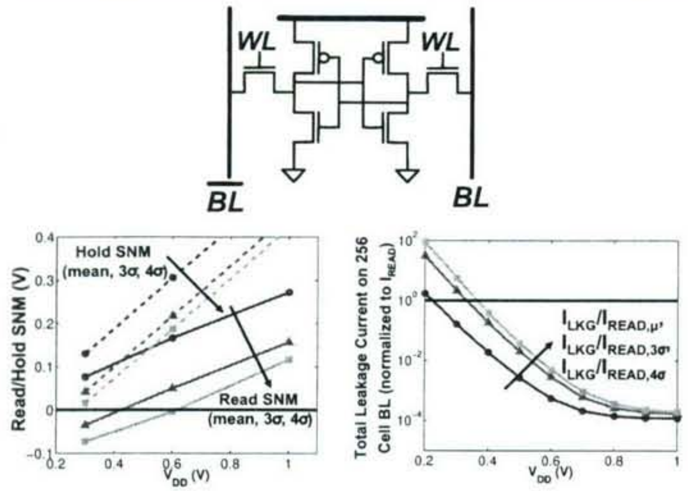


Figure 18.4.1: 6T cell SNM and bitline leakage (normalized to I_{READ}) demonstrating loss of functionality at low voltages.

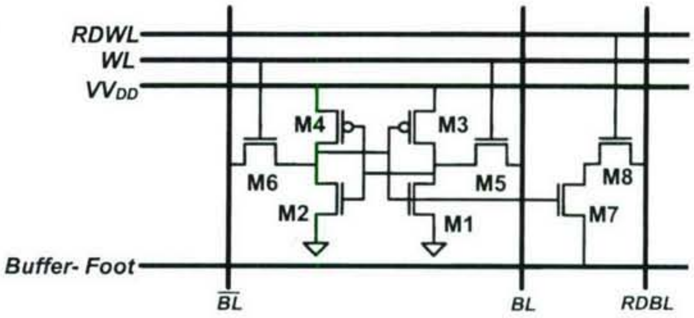


Figure 18.4.2: 8T cell enabling low-voltage read/write and sensing.

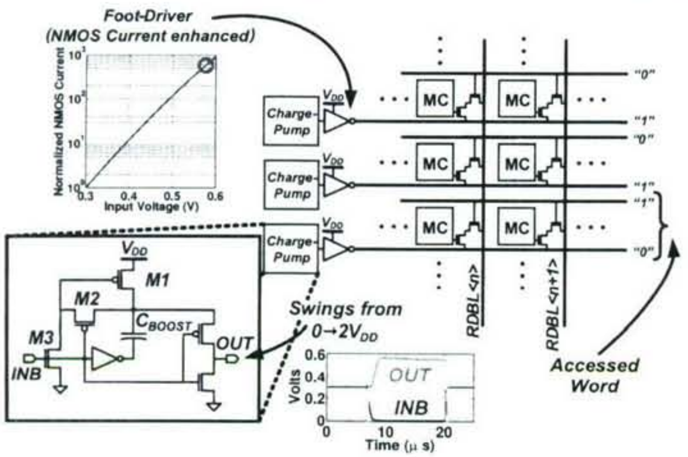


Figure 18.4.3: Circuitry to eliminate sub-V_t leakage from unaccessed read-buffers. Peripheral charge-pumps ensure buffer-foot drivers do not limit I_{READ}.

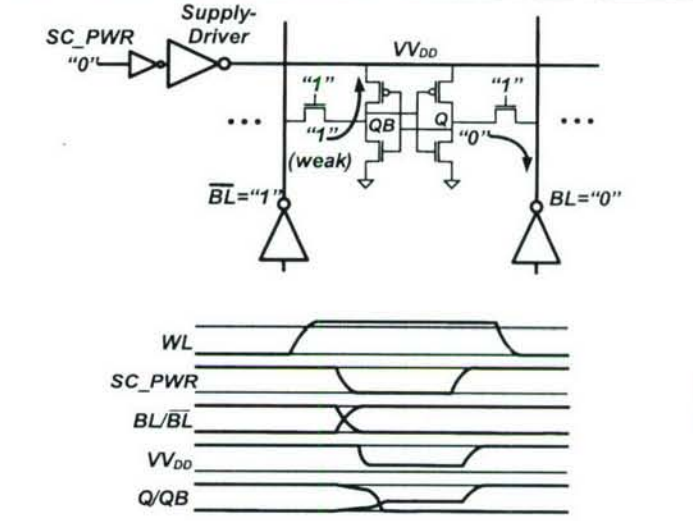


Figure 18.4.4: Cell write performed by weakening local feed-back. Cell supply settles to low intermediate voltage determined by supply driver and write drivers.

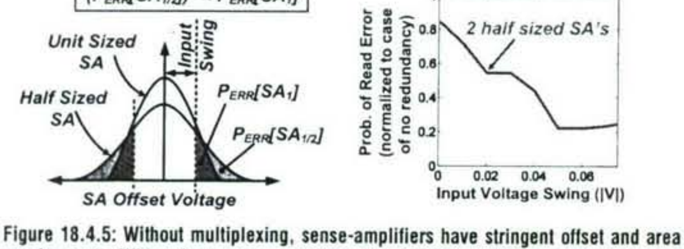


Figure 18.4.5: Without multiplexing, sense-amplifiers have stringent offset and area requirements. With redundancy, errors depend on joint probabilities, improving offset for a given area constraint.

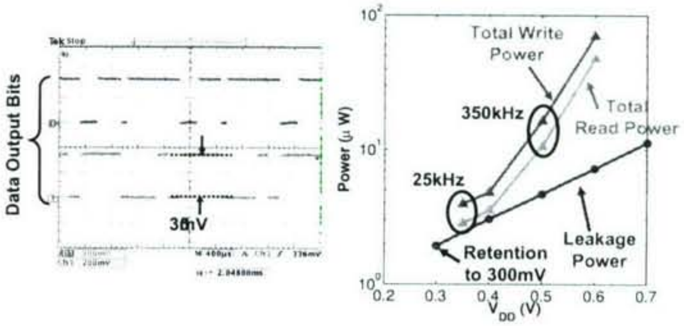


Figure 18.4.6: Scope output and measurements of 65nm test-chip. Array reads and writes at 350mV. Data is correctly retained at 300mV.

Continued on Page 606

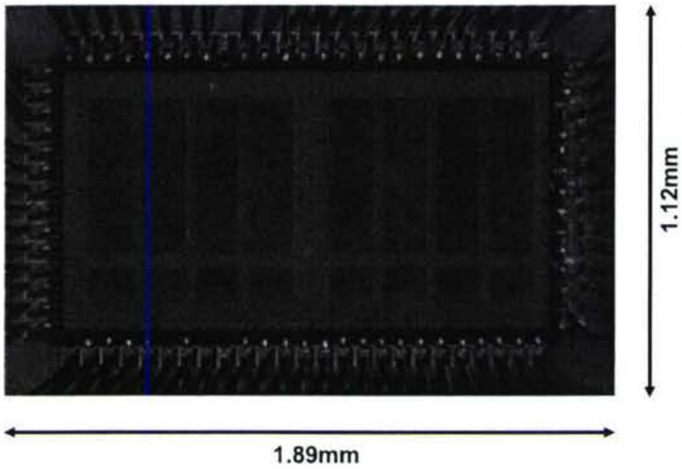


Figure 18.4.7: Die photograph of 256kb 8T SRAM in 65nm CMOS.

DESIGN OF AN ULTRA-LOW-VOLTAGE UWB BASEBAND PROCESSOR

Vivienne Sze, Anantha P. Chandrakasan

Massachusetts Institute of Technology

ABSTRACT

This paper presents an energy-efficient UWB baseband processor that achieves a 100-Mbps throughput while operating at a sub-threshold supply voltage of 0.4 V. While sub-threshold operation is traditionally used for low energy, low performance applications (e.g. wrist-watches), this work examines how it can be applied to low energy, high performance applications using extreme parallelism. Measured results for a 20-pJ/bit 0.4-V UWB baseband processor are presented. Power gating is used to reduce leakage energy.

1. BACKGROUND INFORMATION

This work was performed during the master's degree program from September 2004 to June 2006 at the Massachusetts Institute of Technology in Cambridge, MA, United States. The submission category is "Operational Chip Design".

2. INTRODUCTION

The consumer electronics industry is exploring the use of Ultra-wideband (UWB) communications, a short-range high-data-rate radio technology, to complement longer range radio technologies such as Wi-Fi, WiMAX, and cellular wide area communications. UWB communications can be used to send data from a host device to other devices within the immediate area, eliminating the need for wires and increasing mobility [1]. The use of UWB as a medium for high-data-rate last-meter wireless links requires that UWB radios be integrated onto battery-operated devices such as mobile phones, handheld devices and sensor nodes. Consequently, there is a need for an energy-efficient UWB transceiver. The main contribution of this work is to demonstrate how extreme parallelism in the digital baseband processor allows for

- sub-threshold operation at 0.4 V to lower energy consumed by the baseband processor
- reduced acquisition time to lower energy consumed by other blocks in the receiver

This is the first work to demonstrate the use of sub-threshold operation for a high performance application.

This paper will begin with a description of the UWB specifications and complete receiver architecture. Next, the main functions of the baseband are discussed. This is followed by a description of how parallelism can be used to achieve an energy-efficient baseband processor and an explanation of the design methodology used to implement it. Finally, the measured results and the test setup used to obtain these results are presented.

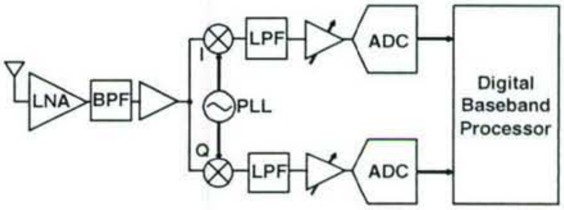


Fig. 1. Block diagram of UWB receiver

3. UWB SPECIFICATIONS AND RECEIVER

The FCC has authorized UWB wireless communications in the 3.1-GHz to 10.6-GHz band with a minimum bandwidth of 500 MHz and a maximum equivalent isotropic radiated power spectral density of -41.3 dBm/MHz [2]. There are two technological approaches for UWB communications: OFDM and pulse-based. This work focuses on the latter using 2-ns binary phase-shift keying (BPSK) pulses.

The receiver, shown in Figure 1, uses a direct-conversion architecture in the front-end and the in-phase and quadrature components are sampled at 500 MSPS by two 5-bit ADCs [3]. For real-time demodulation of the UWB packet, the digital baseband must perform the signal processing with a throughput of 500 MSPS. Synchronization is performed entirely in the digital domain and only the automatic gain control (AGC) is fed back to the analog domain. The baseband was implemented using a standard digital logic cell library in the 90-nm process.

The UWB packets are built from a sequence of BPSK pulses with a 500-MHz bandwidth. The transmitter generates approximate Gaussian pulses and up-converts the packet to one of 14 channels in the 3.1-GHz to 10.6-GHz band. The physical-layer of each packet, shown in Figure 2, can be divided into two sections: preamble and payload. The preamble contains repetitions of a N_c=31 bit Gold code (PN sequence)

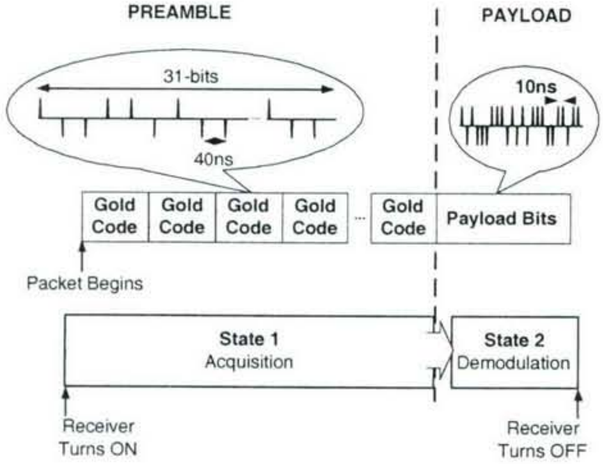


Fig. 2. UWB physical-layer packet format

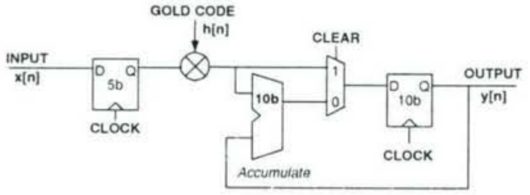


Fig. 3. Correlator Architecture

sent at a pulse repetition frequency (PRF) of 25 MHz. The payload contains the actual data and is sent at a PRF of 100 MHz for a 100-Mbps data rate with no channel coding.

4. UWB BASEBAND PROCESSOR

4.1. UWB Baseband Operation

The baseband processor implements acquisition, synchronization and demodulation by transitioning between two states of operation. The preamble is used by the receiver to achieve acquisition and synchronization. At the receiver, the baseband processor computes the cross-correlation function ($y[n]$) between the incoming noisy preamble ($x[n]$) and a clean template of the 31-bit Gold code sequence ($h[n]$).

$$y[n] = \sum_{k=0}^{N_c-1} x[k] \times h[k-n]$$

The computation shown above is performed with the use of a correlator (Figure 3).

Peak detection is performed on the cross-correlation to achieve signal acquisition as well as synchronization. The cross-correlation also provides the channel estimation. Following synchronization, the baseband performs demodulation of the payload bits. Demodulation involves the use of a 5-fingered RAKE receiver to collect and optimally combine the signal energy received on the multiple echo paths using the tap gains determined by the channel estimation. A hard decision is made at the output of the maximum ratio combiner (MRC) to resolve a bit.

The total energy spent on receiving the UWB signal can be divided into two components: acquisition (preamble) energy and demodulation (payload) energy. One of the goals of this work is to reduce the energy spent by the receiver on acquisition. Since this energy does not go directly towards the demodulation of the data, it is seen as overhead energy. During short bursty traffic, where the payload is small, this overhead energy accounts for a significant portion of the total packet energy. Therefore, it is desirable to minimize the amount of overhead energy per packet. The majority of this overhead energy goes into the computation of the cross-correlation function. In this work, we take two different approaches towards reducing this overhead energy, both of which exploit the use of parallelism.

4.2. Sub-threshold Operation

There are a fixed number of operations required by the baseband in order to compute the cross-correlation function, and therefore in order to reduce the energy of the baseband, we need to reduce its energy per operation. The first approach involves scaling down the supply voltage (V_{DD}) such that the correlator, which computes the cross-correlation, operates at its minimum energy point [4]. The minimum energy point occurs since the total energy per operation is composed of dynamic energy and leakage energy.

$$\begin{aligned} E_{total} &= E_{dynamic} + E_{leakage} \\ &= C_{eff}V_{DD}^2 + I_{leak}V_{DD}T_{delay} \end{aligned}$$

From the above equation we see that lowering V_{DD} decreases the dynamic energy. While reducing V_{DD} reduces the leakage power, it also increases the delay (T_{delay}) of the gates. When the V_{DD} is above the threshold voltage of the device, the delay increases linearly with V_{DD}, and there is no significant change in the leakage energy; however, when V_{DD} drops below the threshold voltage of the device, both delay and leakage energy increase exponentially. Since the dynamic energy and leakage energy scale in opposite directions as V_{DD} decreases, a minimum energy point occurs in the sub-threshold region. *Spectre* simulations performed on the correlator indicate that the minimum energy point occurs at 0.3 V, which gives a 9X energy reduction as compared to the full-scale 1-V operation (Figure 4). Ideally, it would be desirable to scale V_{DD} such that the baseband operates at this minimum energy point.

However, as previously mentioned, the baseband processing must sustain a throughput of 500 MSPS in order to achieve

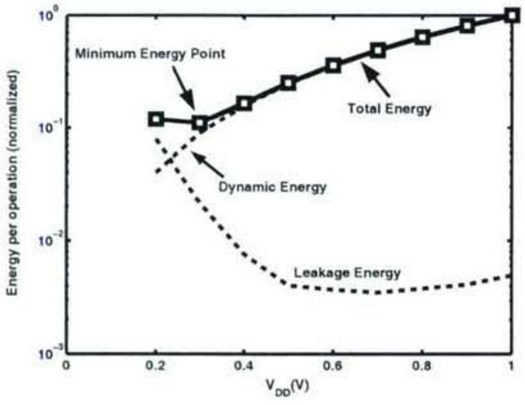


Fig. 4. Simulated energy plot for the correlator.

real-time demodulation. This can be achieved by a single correlator operating at a frequency of 500 MHz with a much higher voltage than 0.3 V, but we have shown that this is not energy efficient. Instead, it is better to operate in subthreshold at a reduced frequency, and utilize parallelism (L) in the baseband to meet the throughput constraint.

In order to refrain from introducing additional complexity due to parallelism, it is preferable that the operating frequency be a factor of the preamble PRF (25 MHz). The operating frequency is equal to 25 MHz if the supply voltage is raised slightly to 0.4 V. Since the minimum energy point is shallow, this slight change in V_{DD} does not cause a significant energy penalty. By operating at 0.4 V rather than 1 V, the energy per operation is reduced by almost 6X. At 25 MHz, the correlators need to be parallelized by a factor of L=20 in order to maintain the 500-MSPS throughput.

This form of parallelism can also be used to reduce the energy spent on the demodulation of the payload bits. The MRC of the RAKE receiver is parallelized by a factor of 4 such that it can operate off the same supply voltage and operating frequency as the rest of the baseband. Combining parallelism with sub-threshold operation delivers energy savings for receiving the entire UWB packet.

4.3. Reduce Acquisition Time

In addition to reducing the energy of the baseband, we would also like to reduce the overhead energy spent by the rest of the blocks in the receiver. The entire receiver must be turned on for the duration of the entire UWB packet. The second approach involves reducing acquisition time to minimize the overall on-time of the receiver, which include the RF front-end, two ADCs and two baseband amplifiers. When combined with duty-cycling, this results in a reduction of the overhead energy.

Reduced acquisition time can be achieved by computing multiple points in the cross-correlation function at the same time. This involves replicating the correlator architecture and operating them in parallel. As the degree of parallelism (M) increases, the maximum time to achieve acquisition decreases by 1/M. The number of Gold code repetitions in the preamble can subsequently be reduced, which results in shorter packets that translates to shorter receiver on-time. Further analysis is presented in [5].

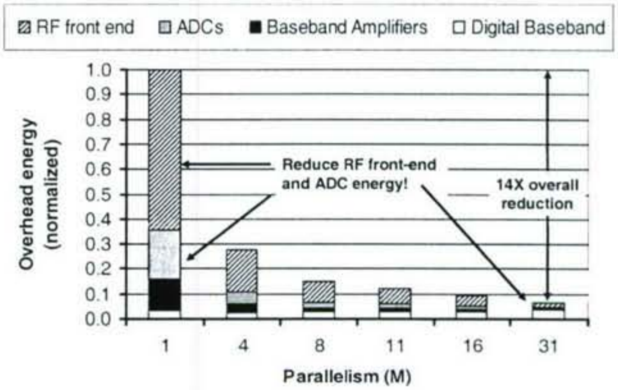


Fig. 5. Breakdown of overhead energy

The impact of this on-time reduction can be seen in the reduction of the overhead energy consumed by the receiver shown in Figure 5. These values were derived from the measured power of the RF front-end, and ADCs, which account for 79% of the total receiver power [3, 6]. When the baseband is parallelized by the length of the Gold code (M=N_C=31), all points of the cross-correlation function can be computed simultaneously, which minimizes the on-time of the receiver, resulting in a 14X reduction in overhead energy. For a 4-kbit packet, this degree of parallelism results in a 43% reduction in the total energy per packet consumed by the receiver.

4.4. Baseband Architecture

The combination of these two approaches results in a highly parallelized implementation with a total of L×M=620 correlators and 4 RAKE MRCs. The parallelized architecture is shown in Figure 6. There are L=20 correlators in each sub-bank in order to maintain the 500-MSPS throughput, and M=31 sub-banks so that all points of the cross-correlation can be computed at the same time. The first form of parallelism, which reduces the energy of the baseband processor, is determined by the frequency of the correlator near its minimum energy point, while the second form, which reduces the energy of the other blocks in the receiver, is dictated by the length of the Gold code sequence (N_c).

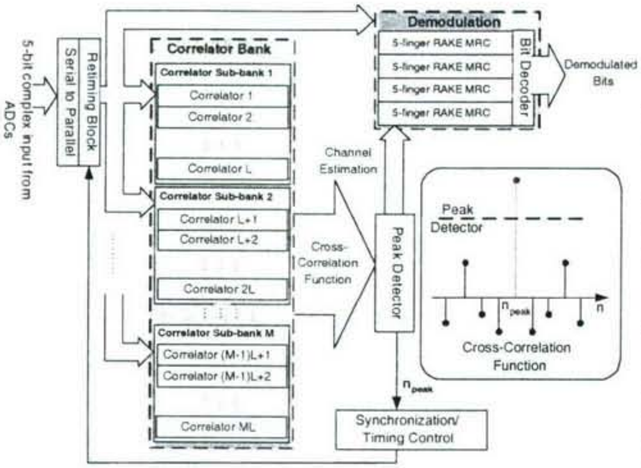


Fig. 6. Architecture of highly parallelized energy efficient UWB baseband processor

5. DESIGN METHODOLOGY

5.1. Circuit Simulation and Implementation Tools

The baseband algorithm was first verified using *MATLAB* to ensure correct functionality. This setup was also useful in generating test vectors. Initially, only the correlator was synthesized by *Synopsys Design Compiler* using STMicroelectronics' 90-nm standard cell library. *Cadence Spectre* was then used to simulate the correlator to determine its minimum energy point. The standard cell library was re-characterized with *Cadence SignalStorm* for the optimum voltage point of 0.4 V. In sub-threshold operation, the delay of the gates decreases with temperature, which is contrary to the behavior in active region operation. This is because I_{off} increases with temperature, while I_{on} decreases with temperature. The corner library characterizations take this into account (i.e. the fast corner used a higher temperature than the slow corner).

With the use of *Perl* scripting, the baseband algorithm was translated into digital circuits written in *Verilog* with the appropriate degree of parallelism (L,M). The entire baseband processor was then synthesized with the 0.4-V library, and *Synopsys Astro* was used for place-and-route. Distributed clock gating was incorporated for further power savings on the clock network. For instance, this ensures that the large correlator bank is not clocked during demodulation.

Also, due to the high degree of parallelism, a hierarchal approach was used to minimize the turn-around time of the EDA tools. Synthesis was performed in the following order:

1. a single correlator
2. the correlator sub-bank (instantiate 20 correlators)
3. the correlator bank (instantiate 31 sub-banks)
4. top-level baseband processor (instantiate correlator bank)

Timing verification that incorporated global variations was performed using *Synopsys PrimeTime*. In sub-threshold operation, the impact of local transistor-to-transistor variations is quite severe. Consequently, for additional variation analysis, Monte Carlo simulations were performed using *Spectre* to verify timing on critical paths. The circuit was also simulated in *Nanosim* with extracted RC parasitics.

5.2. I/O Implementation Considerations

It was desirable to minimize the size of the die since this allows for a smaller package with lower bond wire inductance and cost. Since the baseband processor was pad limited, steps were taken to reduce the number of I/O pads. The control bits were read in serially through a shift register which helped reduce 85 signals to 4. In addition, rather than having 200 pads [=5 (Number of bits in ADC)× 2 (for in-phase and quadrature components)× L (Degree of parallelism for throughput)] allocated to the input signal, the baseband only took in 5 parallel complex inputs at a time rather than 20, which reduced the input signal pads to 50. The penalty for this was that a serial-to-parallel converter was required internally on the baseband processor resulting in a second clock domain of 100 MHz. The timing constraints for signals crossing the clock domains have to be carefully set and verified. Finally, test points at various stages of the baseband were passed through a mux such that they used minimum number of pads without compromising the testability of the chip.

The completed chip had 152 I/O pads. A 144 CQFP package was used, which required 8 ground pads to be down-bonded to the package paddle, and the paddle was bonded to 21 package ground pins.

Since 50 input signals enter the chip at 100 MHz, a surface mount socket was used to minimize inductance. A 4-layer PCB was used in order to have a solid ground plane and to minimize routing of the 100-MHz signals. The board layout is shown in Figure 7.

5.3. Test Equipment and Setup

A Keithly sourcemeter, a Textronix 500 MHz real-time scope, arbitrary waveform generator, logic analyzer and pattern generator were used for testing and chip measurement. The minimum output voltage of the pattern generator was 1.2 V which overdrove the input pads of the chip. The I/O pads operated off a 1-V supply, while the core operated off a separate 0.4-V supply.

6. PERFORMANCE RESULTS

The baseband processor, shown in Figure 7, demonstrates 100-Mbps operation in the sub-threshold region at 0.4 V with

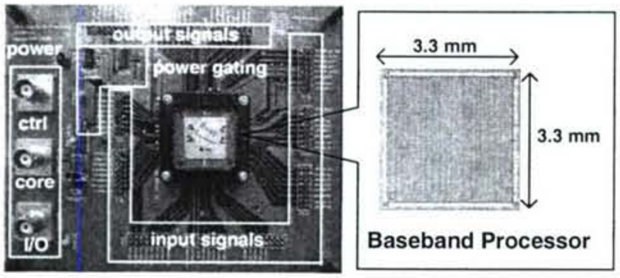


Fig. 7. PCB and die photo of baseband processor

an operating frequency of 25 MHz. A summary of the performance metrics is shown in Table 1. As previously mentioned, we are pad-limited and consequently, only 23% of the die area is active; the rest is filled with decoupling MOS capacitors. The active area of the baseband is comparable to the total active area of the RF front-end and ADC [3, 6].

The breakdown of the energy per bit consumed by the baseband is shown in Figure 8. For a 4-kbit packet, the average energy per bit consumed by the baseband processor is 20 pJ with 3 pJ going towards acquisition and 17 pJ going to demodulation.

Chip Specifications	
Process Technology	90 nm
Die Size	3.3 mm × 3.3 mm
Bit Rate	100 Mbps
Transistor Count	2.6M (2.8k per correlator)
Operating Frequency	25 MHz
Supply Voltage	0.4 V
Power Consumption	
Acquisition	7 mW
Demodulation	1.7 mW

Table 1. Chip Measurements

By operating in the sub-threshold region, the baseband processor achieves significant energy savings as compared with current state-of-the-art UWB baseband transceivers. Under similar packet length conditions, this baseband processor has a reported energy per pulse of less than 1/600 of [7] and 1/5 of [8] (Table 2).

6.1. Power Gating

Both forms of parallelism assume that the receiver can be powered off. Off-chip power gating was used to demonstrate this with the baseband processor (Figure 9). Power gating involves gating the leakage current when the system is idle. A Fairchild NFET was used as the gating transistor.

Realistically, power gating itself costs energy; specifically, the energy required to switch the gating transistor and the recovery energy required to bring the virtual V_{DD} back up to 0.4 V. There is a minimum amount of time that the system must be powered off in order for power gating to be advantageous. This time, known as the break-even time, occurs when the savings in leakage energy is greater than the cost of power gating. Given that the leakage power of the baseband is 745 μW, the break-even time was determined to be 137 μs. A shut-off signal for power gating is automatically generated by the baseband when the packet is completely demodulated. The turn-on signal could be generated at a higher level (e.g. MAC layer).

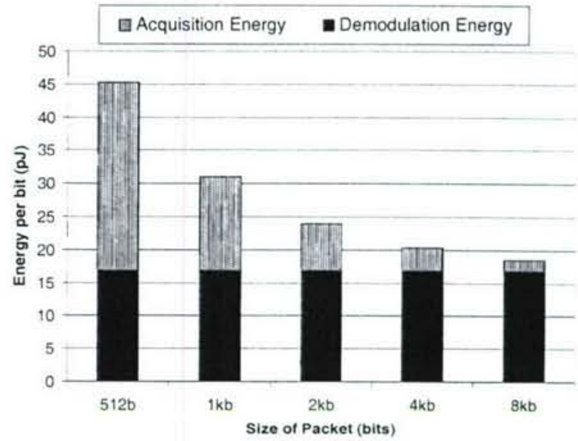


Fig. 8. Breakdown of energy per bit consumed by baseband

Off-chip digital gates were used to implement the power gating control logic (Figure 7). The off-chip gating transistor has a 3-V switching voltage, which required that the control logic operate at 3 V, and a level converter be used to interface the control logic with the 1-V shut-off signal from the baseband processor. A separate 3-V supply voltage was used to power this off-chip control logic.

The instantaneous power of the various states of operation is shown in Figure 10. To obtain this measurement, a 10-Ω resistor was inserted between the sourcemeter and node VDD* labeled in Figure 9 to measure the current. The sourcemeter operated in the 4-wire sense mode in order to maintain a 0.4-V

	[7]	[8]	This Work
Process Technology	0.18 μm	0.18 μm	90 nm
Supply Voltage	1.8 V	1.2 V	0.4 V
Data Rate	193 kbps	62.5 Mbps	100 Mbps
Energy Per Pulse	12.5 nJ	107 pJ	20 pJ

Table 2. Comparison with the state-of-the-art

drop across the gating transistor and the baseband processor.

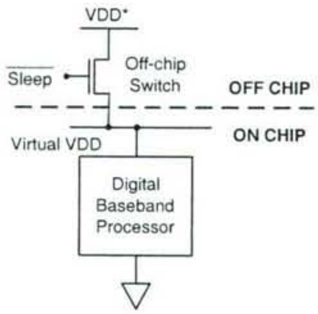


Fig. 9. Off-chip Power Gating

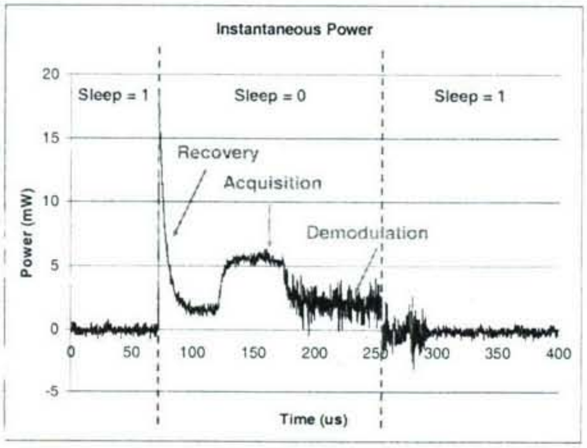


Fig. 10. Measured instantaneous power for various states

7. CONCLUSIONS

Extreme parallelism can be exploited to reduce acquisition time in order to minimize receiver energy, and to enable the use sub-threshold operation for high performance applications. The analysis in this paper can be mapped to other high performance communication applications using sub-threshold operation and parallelism.

8. ACKNOWLEDGEMENTS

The authors would like to thank Raúl Blázquez for technical discussions and STMicroelectronics for providing process and cell library access. This research is sponsored by DARPA and an NSERC Fellowship.

9. REFERENCES

[1] "Ultra-Wideband (UWB) Technology, Technology & Research at Intel," http://www.intel.com/technology/comms/uwb/.

[2] "Ultra-wideband First Report and Order," Tech. Rep., Federal Communications Commission, February 2002.

[3] B. P. Ginsburg and A.P. Chandrakasan, "Dual Scalable 500MS/s, 5b Time-Interleaved SAR ADCs for UWB Applications," in *IEEE Custom Integrated Circuits Conference*, September 2005.

[4] B. H. Calhoun and A. P. Chandrakasan, "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits," in *International Symposium on Low Power Electronics and Design*, August 2004.

[5] V. Sze, R. Blazquez, M. Bhardwaj, and A.P. Chandrakasan, "An Energy Efficient Sub-Threshold Baseband Processor Architecture For Pulsed Ultra-Wideband Communications," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[6] F. S. Lee and A.P. Chandrakasan, "A BiCMOS Ultra-Wideband 3.1-10.6GHz Front-End," in *IEEE Custom Integrated Circuits Conference*, September 2005.

[7] R. Blazquez, P.P. Newaskar, F.S. Lee, and A.P. Chandrakasan, "A Baseband Processor for Impulse Ultra-Wideband Communications," *IEEE Journal Of Solid-State Circuits*, vol. 40, no. 9, pp. 1821–1828, September 2005.

[8] C.-H. Yang, K.-H. Chen, and T.-D. Chiueh, "A 1.2V 6.7mW Impulse-Radio UWB Baseband Transceiver," in *IEEE International Solid-State Circuits Conference*, February 2005.

A Parallel Energy Efficient 100Mbps Ultra-Wideband Radio Baseband

Brian P. Ginsburg, Vivienne Sze, and Anantha P. Chandrakasan
Microsystems Technology Laboratory
Massachusetts Institute of Technology
Cambridge, MA, USA, 02139

Abstract: *An analog-to-digital converter (ADC) and a digital baseband processor for an ultra-wideband (UWB) radio receiver perform sampling and demodulation of 100-Mbps UWB pulses. Parallelism is used to achieve the high throughput with state-of-the-art power consumption. The 5-bit 500-MS/s ADC consumes only 6 mW, and the digital processor operates at 0.4 V.*

Keywords: Parallelism; low-power; ultra-wideband radio; analog-to-digital converter; digital processor.

Introduction

Ultra-wideband (UWB) radio is an emerging technology that shows promise for very-high-data-rate wireless communication over short distances. Applications of UWB include battery-operated devices such as mobile phones, handheld devices and sensor nodes. Consequently, there is a strong demand for an energy efficient UWB system.

The FCC defines UWB signals as having 10-dB bandwidths greater than 500 MHz, and limits transmission power density to less than –41.3 dBm/MHz in the 3.1–10.6-GHz band [1]. The primary technical approaches for high-data-rate UWB communication are OFDM [2] and pulse-based [3] solutions. The baseband presented in this work targets a custom pulse-based radio. BPSK-modulated Gaussian pulses are transmitted at a pulse repetition frequency (PRF) of 100 MHz in one of 14 500-MHz-wide channels within the UWB band [4]. After down-conversion by a direct-conversion RF front-end [5] (Figure 1), the received complex signal occupies 0–250 MHz.

Parallelism is exploited in both the analog-to-digital converter (ADC) and the baseband processor in order to achieve the 100-Mbps throughput with minimum power consumption. In the ADC, time-interleaving allows the use of the energy efficient successive approximation register (SAR) architecture [6], while in the baseband processor, it enables operation using an ultra-low-voltage supply [7].

500-MS/s, 5-bit Analog-to-Digital Converter

The 250-MHz down-converted pulses require a 500-MS/s Nyquist converter, but the required resolution is limited to 4–5 bits [8]. Flash ADCs are the typical choice for this high-speed, low-resolution regime. A flash converter compares the input, in parallel, to every possible threshold voltage and determines the binary output in a single clock cycle. This use of voltage parallelism enables the highest-speed ADC operation, but it requires an exponential growth in the number of comparators with the resolution. This undesirable complexity characteristic has long motivated the choice of other architectures. The successive approximation register (SAR) topology has only a linear growth in the number of comparisons with the resolution; however, it computes each bit of the digital output sequentially and therefore requires multiple clock periods to resolve a conversion, limiting conversion speed. Time-interleaving [9] uses parallel channels, sampling at fixed time-intervals to increase the conversion time of any single channel, permitting use of the energy efficient SAR architecture for this high-speed application [10]. An energy comparison between the flash and time-interleaved SAR architecture is presented in [11].

One limitation to the general use of parallelism is the requirement of independent processing from sample to sample. Successive samples of any true Nyquist converter, however, should be assumed to be completely independent of each other. Thus, processing the samples in parallel should give identical results to processing them serially. In practice, however, mismatches between channels, can negatively impact ADC performance. The three primary mismatch concerns are offset, gain, and timing skew. The design of the time-interleaved SAR ADC is presented below, specifically addressing our solutions to mismatches.

Top-Level Architecture: The SAR algorithm requires one period to decide each of the output bits plus one period for sampling. With six time-interleaved channels, the internal channel clock period matches the overall sampling clock. Thus, only one clock needs to be generated and distributed. Besides easing clock distribution requirements, this also minimizes timing skew between channels. A balanced layout for this single sampling clock is sufficient to reduce errors arising from timing skew to below the 5-bit level. The top-level block diagram of the 6-channel ADC is shown in Figure 2. Synchronization is performed by passing a start token that signals when a channel should begin sampling. This keeps the overhead associated with time-interleaving to a minimum.

Channel Circuits: Each channel is composed of a capacitive DAC, a comparator, and digital control logic, often referred to as the SAR itself. The DAC is the split capacitor array [12], which features decreased switching energy and faster switching speed than the conventional binary weighted capacitor array. Gain mismatch between channels is limited by capacitor matching, and the unit capacitor size is thereby chosen conservatively.

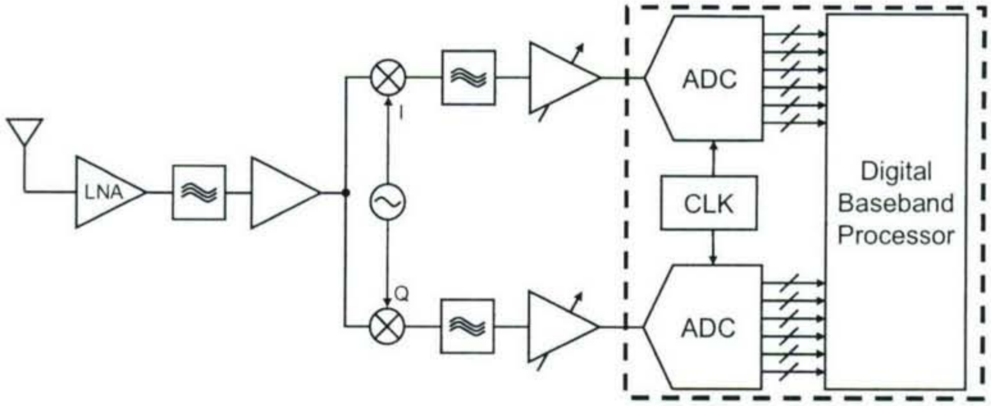


Figure 1. UWB direct conversion receiver block diagram. Baseband is highlighted.

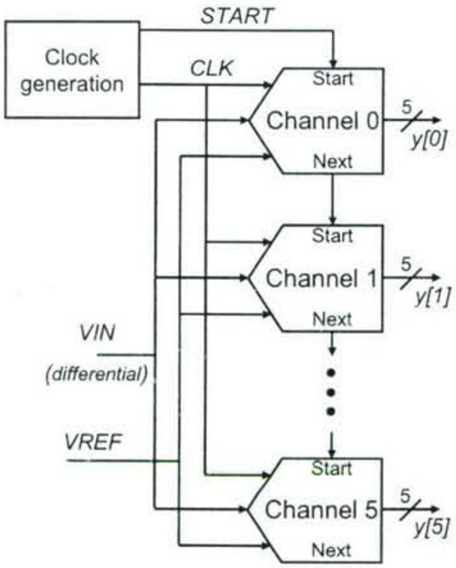


Figure 2. Block diagram of 6-way time-interleaved SAR ADC.

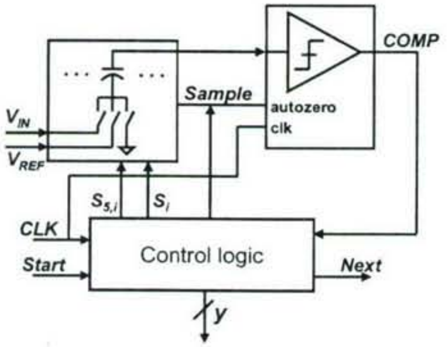


Figure 3. Block diagram of the SAR channel.

The comparator uses a two stage autozeroed preamplifier and a regenerative latch. The preamplifiers reduce the large offset voltage of the latch to below one quarter of the LSB voltage when referred to the input of the entire comparator chain, sufficient to limit offset mismatch. All of the transistors in the comparator have longer than the minimum channel length in order to improve matching and output impedance.

Implementation and Measured Results: The ADC has been fabricated in a 65-nm CMOS process. A photograph of the 1.9 x 1.4 mm die is shown in Figure 4. The input and clock paths use a fully balanced layout in the middle of the die.

The effect of mismatch between channels can be seen in the FFT in Figure 5. Distortion from the measured $0.3V_{LSB}$ offset variation appears as spurs (e)–(f) at multiples of the channel sampling frequency. Spurs (a)–(d) arise from timing and gain errors, dominated by the former in this implementation. The measured gain error is 0.9%. The ADC achieves full Nyquist operation, with the effective number of bits dropping from 4.5 at DC to 4 at Nyquist. The measured 6-mW power consumption is split roughly evenly between the analog and digital supplies. The ADC performance summary is listed in Table 1.

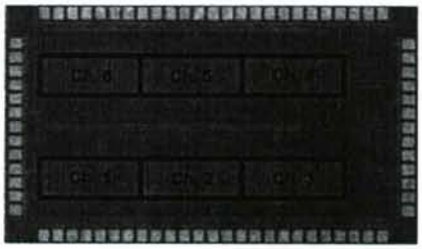


Figure 4. Die photograph of ADC.

Table 1. Summary of ADC Performance

Technology	65-nm CMOS 1P6M
Supply Voltage	1.2 V
Sampling Rate/Resolution	500 MHz/5 bit
SNDR (f_{in} = 239 MHz)	26.1 dB
DNL/INL	0.16/0.26 LSBs
Power (analog/digital)	2.86/3.06 mW
Active Area	0.65 mm x 1.4 mm

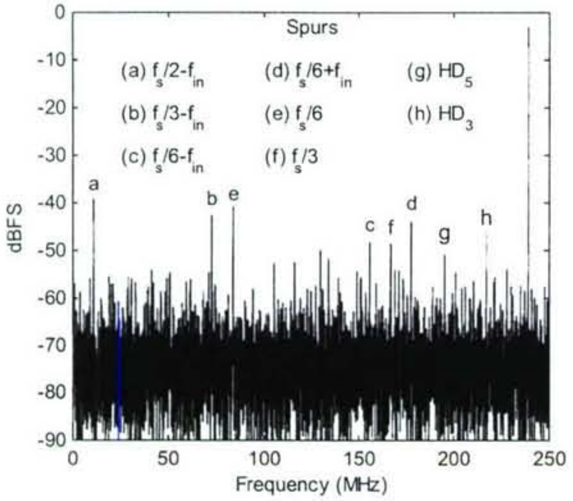


Figure 5. FFT of 239 MHz input with dominant spurs labeled.

UWB Baseband Processor

The digital baseband processor, shown in Figure 6, receives a 500-MS/s signal from the ADC and performs acquisition and demodulation. The packet structure of the received signal is shown in Figure 7. The preamble contains repetitions of a 31-bit PN sequence with a PRF of 25 MHz, while the payload contains the actual data that is sent at a PRF of 100 MHz for a 100-Mbps data rate with no channel coding. During acquisition, a correlator is used to compute the cross-correlation function between the incoming noisy preamble and a clean template of the 31-bit PN sequence. Peak detection is performed on the cross-correlation to achieve signal acquisition. Demodulation is then performed on the payload with a 5-fingered RAKE receiver, and a hard decision is made at the output of the maximum ratio combiner (MRC) to resolve a bit.

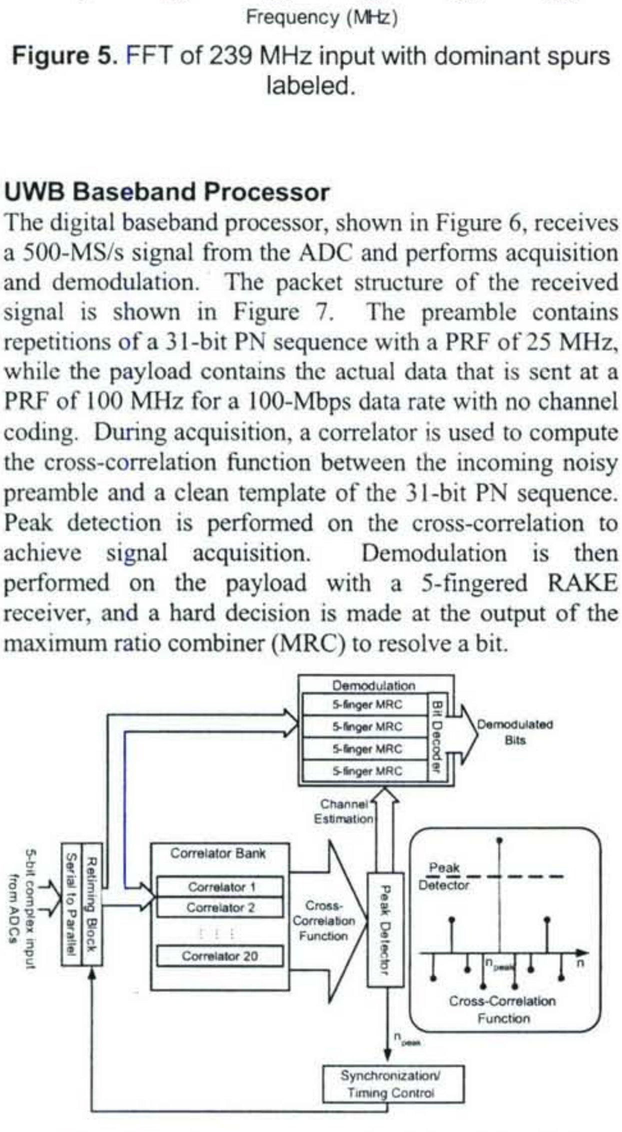


Figure 6. Block diagram of parallelized digital baseband processor

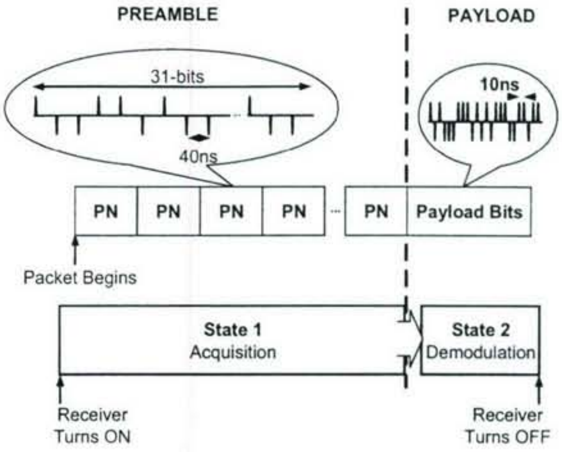


Figure 7. UWB physical-layer packet format

The energy of the baseband processor is reduced by aggressively scaling down its supply voltage (V_{DD}) such that the correlator operates near its minimum energy point [13]. The minimum energy point occurs because the total energy per operation is composed of dynamic energy and leakage energy.

$$E_{total} = E_{dynamic} + E_{leakage} = C_{eff} V_{DD}^{2} + I_{leak} V_{DD} T_{delay}$$

From the above equation, we see that lowering V_{DD} decreases the dynamic energy. While reducing V_{DD} reduces the leakage power, it also increases the delay (T_{delay}) of the gates. When the V_{DD} is above the threshold voltage of the device, the delay increases linearly with V_{DD}, and there is no significant change in the leakage energy; however, when V_{DD} drops below the threshold voltage of the device, both delay and leakage energy increase exponentially. Since the dynamic energy and leakage energy scale in opposite directions as V_{DD} decreases, a minimum energy point occurs in the sub-threshold region.

Simulations performed on the correlator, designed in a standard-V_T 90-nm CMOS process, indicate that the minimum energy point occurs at 0.3 V, which gives a 9x energy reduction as compared to the full-scale 1-V operation (Figure 8) [7]. Ideally, it would be desirable to scale V_{DD} such that the baseband operates at this minimum energy point.

For real-time acquisition and demodulation of a UWB packet, the baseband processor must perform signal processing with a throughput of 500 MS/s. This can be achieved by a single correlator operating at a frequency of 500 MHz with a much higher voltage than 0.3 V, but we have shown that this is not energy efficient. Instead, it is better to operate at a lower voltage with a reduced frequency, and utilize parallelism in the baseband processor to meet the throughput constraint.

In order to refrain from introducing additional complexity due to parallelism, it is preferable that the operating

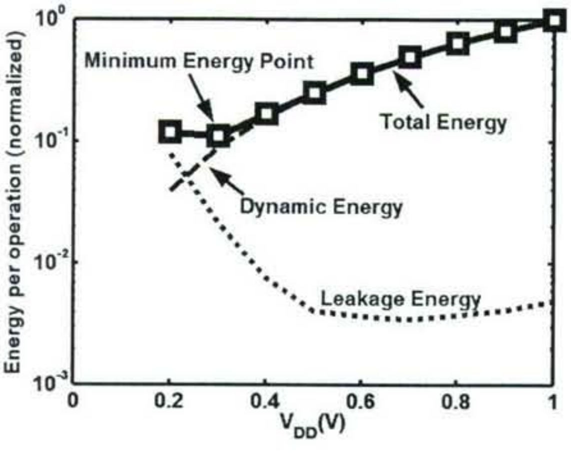


Figure 8. Simulated energy plot for the correlator

frequency be a factor of the preamble's PRF (25 MHz) such that an integer number of pulses are processed per clock cycle. The operating frequency is equal to 25 MHz if the supply voltage is raised slightly to 0.4 V. Since the minimum energy point is shallow, this slight change in V_{DD} does not cause a significant energy penalty. By operating at 0.4 V rather than 1 V, the energy per operation is reduced by almost 6x. At 25 MHz, the correlators need to be parallelized by a factor of 20 in order to maintain the 500-MS/s throughput.

In addition, the MRC of the RAKE receiver is parallelized by a factor of 4 such that it can operate off the same supply voltage and operating frequency as the rest of the baseband processor. This also reduces the energy required for demodulation. Combining parallelism with ultra-low-voltage operation delivers energy savings for receiving the entire UWB packet.

Conclusion

An energy-efficient baseband for a UWB radio has been presented. Parallelism has enabled the very low power consumption in this high performance application. The low complexity but high latency successive approximation register ADC architecture is combined with time-interleaving to achieve the desired throughput and performance in deep-submicron CMOS. The highly parallelized packet acquisition leads to a significant reduction in operating voltage. The baseband presented here can be integrated in a single-chip solution in a highly energy-efficient manner by increasing the number of time-interleaved ADC channels. With 20 channels, each channel would directly feed one set of correlator banks, and the channels themselves could translate the increased conversion into reduced operating voltages for further energy savings.

Acknowledgments

This work is funded by DARPA, an NDSEG Fellowship, and an NSERC Fellowship. The authors would like to thank Texas Instruments and STMicroelectronics for fabrication services.

References

[1] Federal Communications Commission, "Ultra-wideband first report and order," FCC 02-48, Feb. 2002.

[2] A. Batra, *et al.*, "Multi-band OFDM physical layer proposal for IEEE 802.15 Task Group 3a," IEEE P802.15-04/0493r0, Sept. 2004.

[3] R. Fisher, *et al.*, "DS-UWB physical layer submission to 802.15 Task Group 3a." IEEE P802.15-04/0137r3., July 2004.

[4] D. D. Wentzloff, *et al.*, "System design considerations for ultra-wideband communication," *IEEE Commun. Mag.*, vol. 43, no. 8, pp. 114 –121, Aug. 2005.

[5] F. S. Lee and A. P. Chandrakasan, "A BiCMOS ultra-wideband 3.1–10.6-GHz front-end," *IEEE J. Solid-State Circuits*, vol. 41, no. 8, pp. 1784–1791, Aug. 2006.

[6] B. P. Ginsburg and A. P. Chandrakasan, "A 500MS/s 5b ADC in 65nm CMOS," in *Symp. on VLSI Circuits Dig. of Tech. Papers*, June 2006, pp. 174–175.

[7] V. Sze, *et al.*, "An energy efficient sub-threshold baseband processor architecture for pulsed ultra-wideband communications," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, May 2006, pp. (III) 908–911.

[8] P. P. Newaskar, R. Blazquez, and A. P. Chandrakasan, "A/D precision requirements for an ultra-wideband radio receiver," in *IEEE Workshop on Signal Processing Systems*, Oct. 2002, pp. 270–275.

[9] W. Black and D. Hodges, "Time interleaved converter arrays," *IEEE J. Solid-State Circuits*, vol. 15, no. 6, pp. 929–938, Dec. 1980.

[10] D. Draxelmayr, "A 6b 600MHz 10mW ADC array in digital 90nm CMOS," in *ISSCC Dig. Tech. Papers*, Feb. 2004, pp. 264–265.

[11] B. P. Ginsburg and A. P. Chandrakasan, "Dual time-interleaved successive approximation register ADCs for an ultra-wideband receiver," *IEEE J. Solid-State Circuits*, vol. 42, Feb. 2007, to be published.

[12] B. P. Ginsburg and A. P. Chandrakasan, "An energy-efficient charge recycling approach for a SAR converter with capacitive DAC," in *Proc. of the IEEE Int. Symp. on Circuits and Systems*, vol. 1, May 2005, pp. 184–187.

[13] B. Calhoun, A. Wang and A. P. Chandrakasan, "Modeling and sizing for minimum energy operation in sub-threshold circuits," in *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1778-1786, September 2005.

Journal Papers

Static Noise Margin Variation for Sub-threshold SRAM in 65-nm CMOS

Benton H. Calhoun, *Member, IEEE*, and Anantha P. Chandrakasan, *Fellow, IEEE*

***Abstract*—The increased importance of lowering power in memory design has produced a trend of operating memories at lower supply voltages. Recent explorations into sub-threshold operation for logic show that minimum energy operation is possible in this region. These two trends suggest a meeting point for energy-constrained applications in which SRAM operates at sub-threshold voltages compatible with the logic. Since sub-threshold voltages leave less room for large static noise margin (SNM), a thorough understanding of the impact of various design decisions and other parameters becomes critical. This paper analyzes SNM for sub-threshold bitcells in a 65-nm process for its dependency on sizing, V_{DD}, temperature, and local and global threshold variation. The V_T variation has the greatest impact on SNM, so we provide a model that allows estimation of the SNM along the worst-case tail of the distribution.**

***Index Terms*—Sub-threshold, sub-threshold memory, SRAM, static noise margin, process variation, voltage scaling.**

I. Introduction

SUB-THRESHOLD digital circuit design has emerged as a low energy solution for applications with strict energy constraints. Analysis of sub-threshold designs has focused on logic circuits (e.g., [1]). SRAMs comprise a significant percentage of the total area for many digital chips as well as the total power [2], [3]. For this reason, SRAM leakage can dominate the total leakage of the chip, and large switched capacitances in the bitlines and wordlines make SRAM accesses costly in terms of energy. Pushing SRAM operation into the sub-threshold region reduces both leakage power and access energy. Also, for system integration, SRAM must become capable of operating at sub-threshold voltages that are compatible with sub-threshold combinational logic. Recent low power memories show a trend of lower voltages with some designs holding state on the edge of the sub-threshold region (e.g., [4]). This scaling promises to continue, leading to sub-threshold storage modes and even sub-threshold operation for SRAMs operating in tandem with sub-threshold logic.

When the bitcell is holding data, its wordline is low so the nMOS access transistors are off. In order to hold its data properly, the back-to-back inverters must maintain bi-stable operating points. The best measure of the ability of these inverters to maintain their state is the bitcell's static noise margin (SNM) [5]. The SNM is the maximum amount of voltage noise that can be introduced at the outputs of the two inverters such that the cell retains its data. SNM quantifies the amount of voltage noise required at the internal nodes of a bitcell to flip the cell's contents.

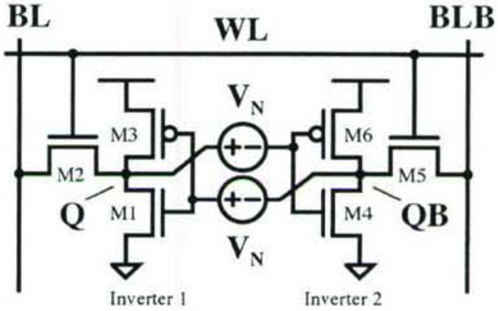


Fig. 1. Schematic for 6T bitcell showing voltage noise sources for finding SNM.

Fig. 1 shows a conceptual setup for modeling SNM [5]. Noise sources having value V_N are introduced at each of the internal nodes in the bitcell. As V_N increases, the stability of the cell changes. Fig. 2 shows the most common way of representing the SNM graphically for a bitcell holding data. The figure plots the voltage transfer characteristic (VTC) of Inverter 2 from Fig. 1 and the inverse VTC from Inverter 1. The resulting two-lobed curve is called a "butterfly curve" and is used to determine the SNM. The SNM is defined as the length of the side of the largest square that can be embedded inside the lobes of the butterfly curve [5]. To understand why this definition holds, consider the case when the value of V_N increases from 0. On the plot, this causes the VTC^{-1} for Inverter 1 in the figure to move downward and the VTC for Inverter 2 to move to the right. Once they both move by the SNM value, the curves meet at only two points. Any further noise flips the cell.

Although the SNM is certainly important during hold, cell stability during active operation represents a more significant limitation to SRAM operation. Specifically, at the onset of a read access, the wordline is "1" and the bitlines are still precharged to "1" as Fig. 3 illustrates. The internal node of the bitcell that represents a zero gets pulled upward through the access transistor due to the voltage dividing effect across the access transistor (M_2, M_5) and drive transistor (M_1, M_4). This increase in voltage severely degrades the SNM during the read operation (read SNM). Fig. 4 shows example butterfly curves during hold and read that illustrate the degradation in SNM during read.

Manuscript received November 4, 2005; revised January 18, 2006. This work was supported by the Defense Advanced Research Projects Agency (DARPA) and by Texas Instruments Incorporated.

B. Calhoun is with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: bcalhoun@virginia.edu).

A. Chandrakasan is with Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: anantha@mtl.mit.edu).

Digital Object Identifier 10.1109/JSSC.2006.873215

0018-9200/$20.00 © 2006 IEEE

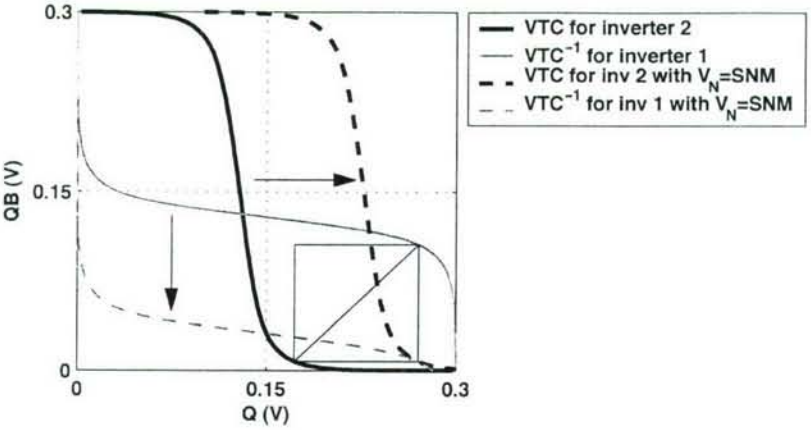


Fig. 2. The length of the side of the largest embedded square in the butterfly curve is the SNM. When both curves move by more than this amount (e.g., V_N = SNM), then the bitcell is mono-stable, losing its data.)

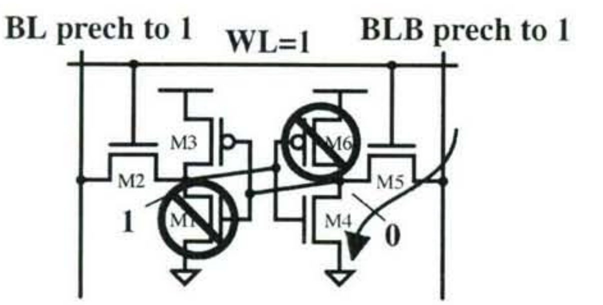


Fig. 3. Schematic of the 6T bitcell at the onset of a read access. WL has just gone high, and both BLs are precharged to $V_{\rm DD}$. The voltage dividing effect across M_4 and M_5 pulls up node Q_B, which should be 0 V, and degrades the SNM.

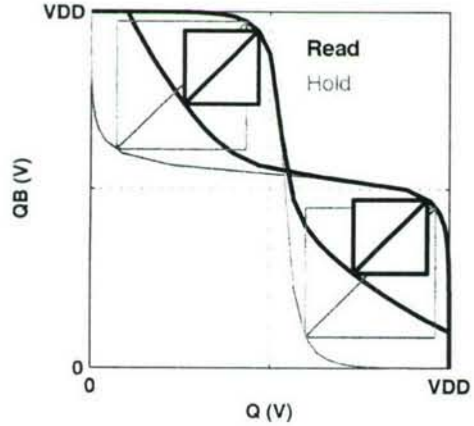


Fig. 4. Example butterfly curve plots for SNM during hold and read.

II. Static Noise Margin

This section evaluates the SNM of six-transistor (6T) SRAM bitcells operating in sub-threshold. We analyze the dependence of SNM during both hold and read modes on supply voltage, temperature, transistor sizes, local transistor mismatch due to random doping variation, and global process variation in a commercial 65-nm technology. We analyze the statistical distribution of SNM with process variation and provide a model for the tail of the probability density function (PDF) that dominates SNM failures [6].

The minimum voltage for retaining bistability was theorized in [7] and modeled for SRAM in [8], but degraded SNM can limit voltage scaling for SRAM designs above this minimum voltage. SNM quantifies the amount of voltage noise required at the internal nodes of a bitcell to flip the cell's contents.

An expression for above-threshold SNM based on long-channel models is given in [5], and [9] models above-threshold SNM for modern processes with process variation. This section builds on previous work by examining SNM for sub-threshold SRAM [6].

A. Modeling Sub-Threshold Static Noise Margin

Lowering $V_{\rm DD}$ reduces gate current much more rapidly than sub-threshold current, so total current in the sub-threshold region can be modeled to first order as

$$I_D = I_S \exp\left(\frac{V_{GS} - V_T}{nV_{th}}\right)\left(1 - \exp\left(\frac{-V_{DS}}{V_{th}}\right)\right). \quad (1)$$

The sub-threshold factor $n = 1 + C_d/C_{ox}$, $V_{th} = kT/q$, and I_S is the current when V_{GS} equals V_T. For simplicity, we treat pMOS parameters as positive values. For the 65-nm technology used in this section, the nMOS drive current is higher in above-threshold than the pMOS for iso-width, but the pMOS current is higher in sub-threshold due to its lower V_T. During hold mode, the wordline is low so M_2 and M_5 have $V_{GS} \leq 0$ and thus negligible current. We can model the cell VTCs ($V_{\rm OUT} = f_{\rm VTC}(V_{\rm IN})$) as those of a simple inverter in sub-threshold.

$$V_{QB} = V_{th}\frac{n_1 n_3}{n_1+n_3}\left(\ln\frac{I_{S3}}{I_{S1}} + \ln\left(\frac{1-\exp\left(\frac{(-V_{\rm DD}+V_Q)}{V_{th}}\right)}{1-\exp\left(-\frac{V_Q}{V_{th}}\right)}\right)\right) + \frac{n_1 V_{\rm DD}}{n_1+n_3} + \frac{n_1 n_3}{n_1+n_3}\left(\frac{V_{T1}}{n_1} - \frac{V_{T3}}{n_3}\right). \quad (2)$$

Referring to Fig. 2, (2) [7] gives the inverse VTC for inverter 1 ($V_{\rm IN} = f^{-1}_{\rm VTC}(V_{\rm OUT})$). The inverse of (2) is given in [10] for

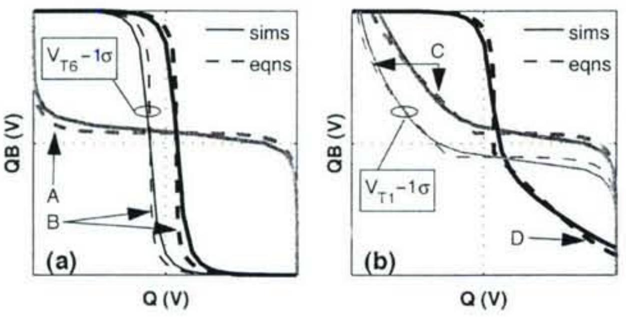


Fig. 5. First-order VTC equations versus simulation. Line A is (2), line B is (3), line C is a piecewise combination of (5) and (2), and line D is a piecewise combination of (3) and the graphical inverse of (5).

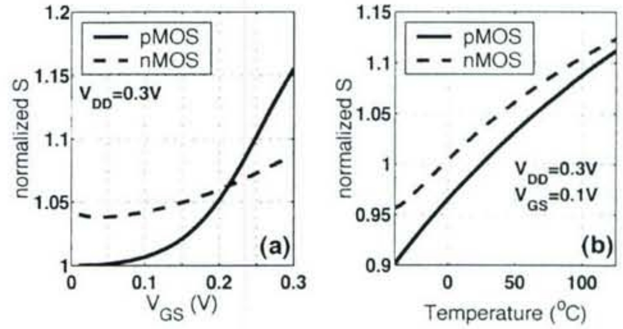


Fig. 6. Changes in sub-threshold slope (S) versus (a) V_{GS} and (b) temperature.

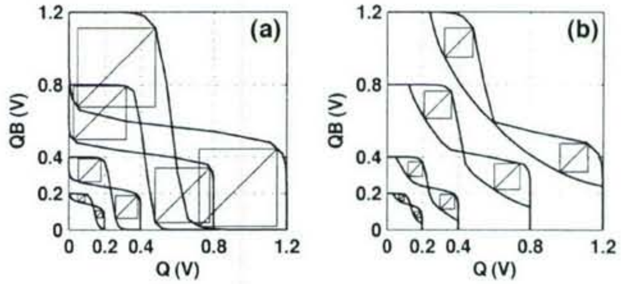


Fig. 7. VTCs for (a) hold and (b) read with varying $V_{\rm DD}$.

matched pMOS and nMOS (same n, V_T, I_S). We give a full solution for $V_{\rm OUT} = f_{\rm VTC}(V_{\rm IN})$ for inverter 2 in (3):

$$V_{QB} = V_{\rm DD} + V_{th} \ln\left(\frac{1-G+\sqrt{(G-1)^2+4\exp\left(\frac{-V_{\rm DD}}{V_{th}}\right)G}}{2}\right) \tag{3}$$

$$G = \exp\left(\frac{n_4+n_6}{n_4 n_6 V_{th}} V_Q - \ln\frac{I_{S6}}{I_{S4}} - \frac{V_{\rm DD}}{n_6 V_{th}} - \frac{1}{V_{th}}\left(\frac{V_{T4}}{n_4} - \frac{V_{T6}}{n_6}\right)\right). \tag{4}$$

Fig. 5(a) plots (2) and (3) against simulation curves for no local mismatch and for 1σ V_T mismatch in M_6.

During a read access, the wordline goes high and the bitlines are precharged to $V_{\rm DD}$ so, if $V_Q = 0$ prior to access, M_1 and M_2 are both on. This creates a voltage division that raises the voltage at Q. Assuming pMOS current is negligible in the region of interest, (5) shows the inverse VTC equation during a read operation near the SNM [4] for inverter 1:

$$V_{QB} = n_1 V_{th} \ln\frac{I_{S2}}{I_{S1}} + n_1 V_{th} \ln\left(\frac{1-\exp\left(\frac{(-V_{\rm DD}+V_Q)}{V_{th}}\right)}{1-\exp\left(-\frac{V_Q}{V_{th}}\right)}\right) + V_{T1} + \frac{n_1}{n_2}(V_{\rm DD} - V_{T2} - V_Q). \tag{5}$$

This equation cannot be inverted analytically, and it applies only to the region of the VTC where $V_{\rm OUT}$ is low. Fig. 5(b) shows (5) and its graphical inverse combined piecewise with (2) and (3) and plotted against simulation for no local mismatch and for 1σ V_T mismatch in M_1 for minimum device sizes at 25 °C.

Graphical or numerical solutions for SNM are easily derived from the VTC equations, although no direct analytical solution exists. The equations provide a good estimate of the behavior of the SNM based on key parameters. One shortcoming of (2)–(5) is the assumption that sub-threshold slope ($S = nV_{th} \ln 10$) is constant for each transistor. Fig. 6(a) shows that S varies with V_{GS}, and Fig. 6(b) shows S changing with temperature without the expected constant slope due to V_{th}. A more crucial problem with (2)–(5) is the assumption that certain currents are negligible. These assumptions break down under certain combinations of V_T variation, rendering the first-order equations inaccurate.

B. Sub-Threshold SNM Dependencies

With embedded SRAM often providing multiple megabits of storage, the SNM of the nominal bitcell becomes largely irrelevant. Variations in processing and in the chip's environment create a distribution of SNM across the bitcells in a given memory, and the worst-case tail of this distribution determines the yield. This section examines the impact of different parameters on SNM in sub-threshold and offers a model for estimating the tail of the SNM density function for process variation.

SNM for a bitcell with ideal VTCs is still limited to $V_{\rm DD}/2$ because of the two sides of the butterfly curve. An upper limit on the change in SNM with $V_{\rm DD}$ is thus 1/2. Fig. 7 shows example butterfly curves at different supply voltages from 1.2 V to 200 mV for both hold and read. Fig. 8 plots SNM versus $V_{\rm DD}$ directly for both hold and read mode. The slopes of the curves confirm that less than 1/2 of $V_{\rm DD}$ noise will translate into SNM changes.

The impact of temperature on SNM in sub-threshold is also not large. Fig. 9 shows SNM versus temperature in sub-threshold and again for strong inversion. The sensitivity in sub-threshold is lower, and varying temperature from −40 °C to 125 °C only alters Read and Hold SNM by 21 mV and 6 mV, respectively. Higher temperatures lower SNM in sub-threshold due to the degraded gain in the inverters that results from worse sub-threshold slope (see Fig. 6(b)). Also, pMOS devices weaken relative to nMOS at higher temperature. Fig. 10 provides example butterfly plots for 0 °C and 100 °C at 1.2 V and 0.3 V.

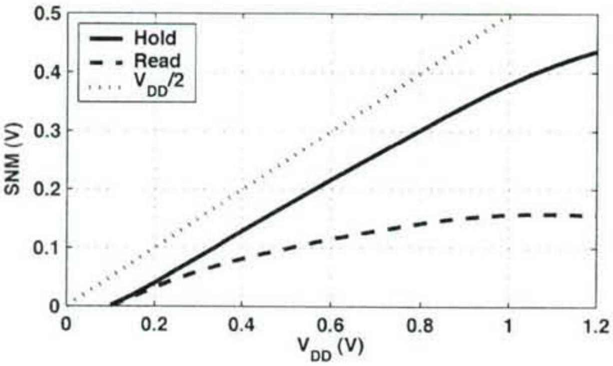


Fig. 8. SNM versus V_{DD}.

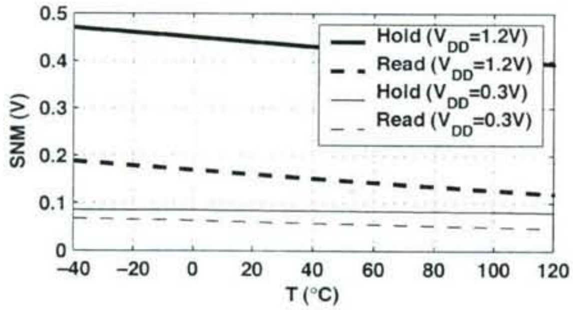


Fig. 9. SNM versus temperature.

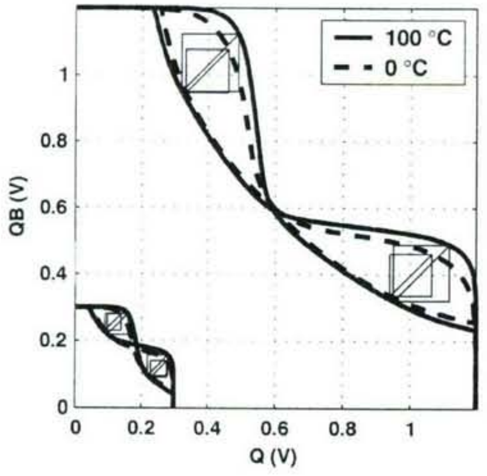


Fig. 10. VTCs during a read access across temperature.

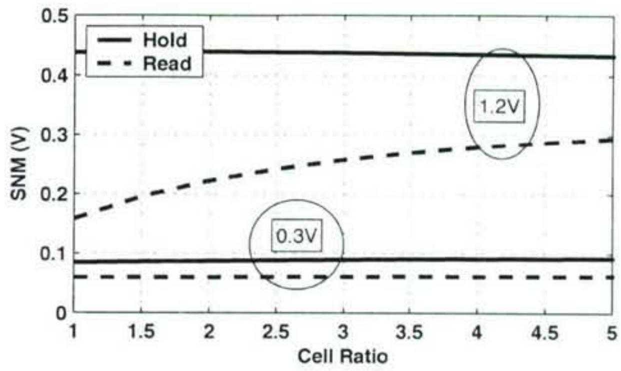


Fig. 11. Cell ratio affects SNM less in sub-threshold.

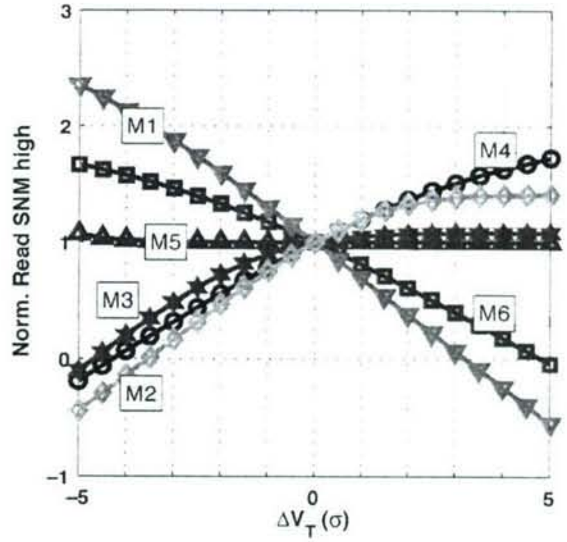


Fig. 12. Dependence of SNM high on single FETs is nearly linear.

In contrast to above-threshold [11], Fig. 11 shows that cell ratio $((W/L)_1/(W/L)_2$ or $(W/L)_4/(W/L)_5)$ has very little impact on SNM during sub-threshold read. In fact, sub-threshold SNM sensitivity to any sizing changes is reduced. The lower impact of sizing is intuitively reasonable considering the exponential dependence of sub-threshold current on other parameters. Mathematically, we can see from (2)–(5) that sizing changes affect I_{Si} linearly and only have a logarithmic impact on the VTCs. One point of caution here is that V_T for deep-submicron devices tends to vary with size as a result of narrow or short channel effects. The impact of this V_T change that might accompany a sizing change is more pronounced. These effects depend on the technology and make general SNM modeling more complicated.

C. Dependence on Random Doping Variation

The randomness of the number of doping atoms and their placement in a MOSFET channel causes random mismatch even in transistors with identical layout [12]. The impact on threshold voltage, whose σ is proportional to $(WL)^{-(1/2)}$, is the worst for minimum sized devices which are common in SRAM. Local variation is a huge problem for SRAM functionality, and it is the subject of many papers (e.g., [13], [14]). The exponential dependence of current on V_T in sub-threshold operation makes this random variation even more influential. Furthermore, the large number of bitcells in many SRAMs makes the tails $(5\sigma$–$6\sigma)$ of the PDF more critical for modeling since the extreme cases are the limiting factor for yield. Previous work has shown that above-threshold SNM is nearly linear with V_T, and modeling its slope as constant allows an approximation of the joint PDF for SNM [9]. Likewise, the sensitivity of above-threshold SNM to V_T is linearized for each transistor in [15].

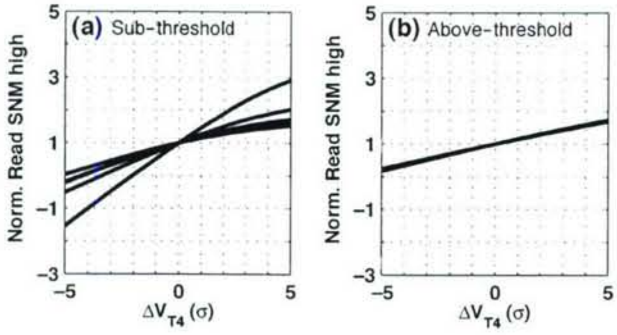


Fig. 13. Dependence of SNM high on a single FET depends on other V_Ts in (a) sub-threshold, unlike for (b) above-threshold.

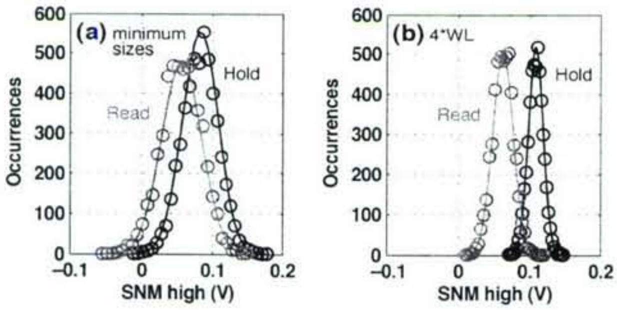


Fig. 14. SNM high and low (not shown) for (a) a minimum sized cell and for (b) $4 * WL$ is normally distributed with random V_T mismatch in all transistors.

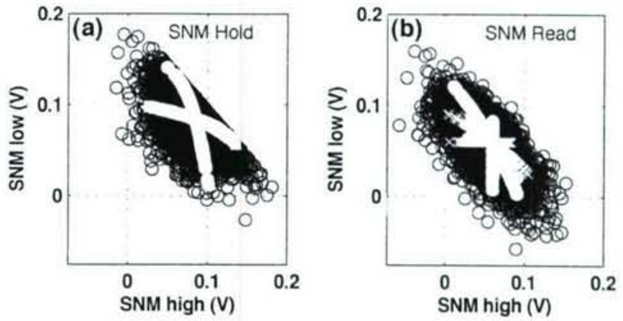


Fig. 15. Scatter plots for SNM high versus SNM low with single FET dependencies overlaid in white.

Fig. 12 shows that, like in strong inversion, the sensitivity of SNM high (the upper-left box in Fig. 4) is nearly linear with each individual V_T. However, Fig. 13(a) shows the relationship between SNM and V_{T4} for a few different random values of the other V_Ts. The obvious dependence of the slope on the other V_Ts prevents using a model of the form $\mathrm{SNM} = \mathrm{SNM}_0 + \sum c_i V_{Ti}$ for sub-threshold SNM. The same is not true of above-threshold, shown in Fig. 13(b), for which a first order series model works well [9], [15].

Fig. 14 shows the results of 5 k-point Monte Carlo (M-C) simulations with random independent V_T mismatch in all transistors. These histograms confirm that sub-threshold SNM at the upper lobe of the butterfly curve (SNM high) is normally distributed. The solid lines show a fitted Gaussian PDF, and the markers show simulation results. Larger sizes for the bitcell clearly have the advertised effect of lowering the variance of V_T as seen in Fig. 14(b). The SNM low PDFs are very similar. The scatter plot in Fig. 15 shows that SNM high and SNM low are correlated. The dependencies for mismatch in each single transistor are overlaid in white for reference. The Hold SNM shows a saturation effect along the upper edge. SNM high and SNM low are not independent because any change to a VTC that increases the SNM at one side tends to decrease SNM at the other side.

The actual SNM that matters for a bitcell is the minimum of SNM high and SNM low. Thus, the random variable $X_{\mathrm{SNM}} = \min(X_{\mathrm{SNMhigh}}, X_{\mathrm{SNMlow}})$. Order statistics can provide us with the PDF for the minimum of n independent, identically distributed (*iid*) random variables, X_i. If f is the PDF, and F is the Cumulative Distribution Function (CDF) for X_i, the PDF of the minimum of two *iid* variables is given in (6):

$$f(\min(X_1, X_2)) = 2f_X(1 - F_X). \tag{6}$$

Although SNM high and SNM low are normally distributed with approximately the same mean and variance, we have previously shown that they are not independent. However, we are less interested in modeling the entire PDF for SNM than we are in modeling the worst-case tail. As previously stated, the tail toward lower SNM is the limiting factor. Let us assume that they are *iid*. Then we can solve for the PDF as

$$f_{\mathrm{SNM}} = 2f_{\mathrm{SNMhigh}}(1 - F_{\mathrm{SNMhigh}}) \tag{7}$$

and the CDF is simply

$$F_{\mathrm{SNM}} = 2F_{\mathrm{SNMhigh}} - (F_{\mathrm{SNMhigh}})^2. \tag{8}$$

Fig. 16 shows the histogram for a 5 k-point M-C simulation of Read SNM plotted on linear axes (a) and semilog axes (b). Clearly, SNM is not normally distributed, and its mean is lower than the mean of SNM high and SNM low. Fig. 16(b) shows that a Gaussian PDF does not match the worst-case tail on the left side of the PDF. On the other hand, the PDF based on (7) provides a good estimate of the worst-case tail. The plot shows that the model does not fit the distribution above the mean. This shortcoming results from the correlation between SNM high and SNM low. Since these two random variables are not *iid*, we cannot claim that the minimum model will always match the tail. However, we can show experimentally that it does offer a good estimate. Thus, the model is a useful tool for evaluating SNM under different design decisions and conditions. This PDF gives the powerful option of estimating the SNM at the worst-case end of the PDF without using extremely long M-C simulations until the design space is narrowed sufficiently.

Fig. 17 shows several estimated PDFs using (7) that are based on data sets of different lengths. These estimates are plotted over a 50 k-point M-C simulation. A 1000-point M-C simulation gives a modeled distribution that overlays the modeled distribution from the 50 k-point case on the plot (<3% error). Using

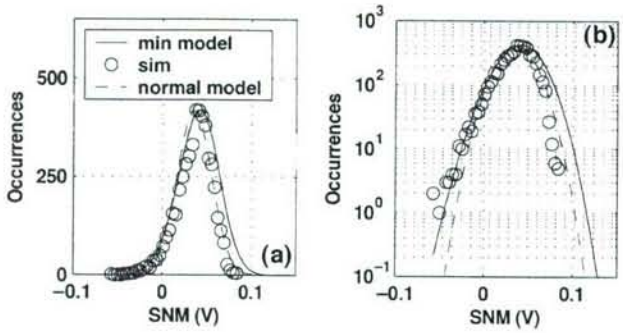


Fig. 16. (a) Histogram of SNM Monte Carlo simulation (circles) with normal PDF (dash) and PDF based on (7) (solid) over-laid. The semilog plot (b) shows that the PDF based on (7) matches the worst-case tail quite well.

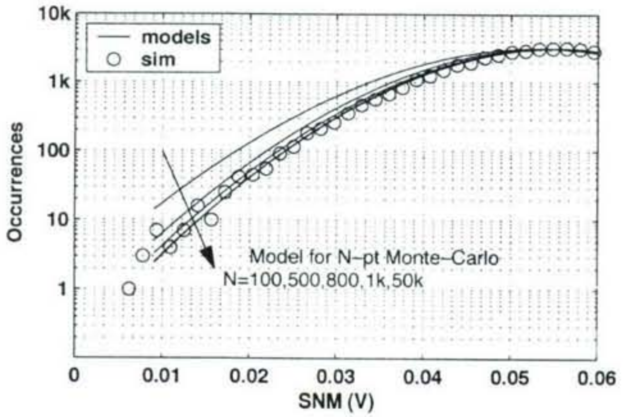


Fig. 17. 50 k-point Monte Carlo simulation for SNM with $4*W/L$ sized transistors. Model based on 1 k-point Monte Carlo data matches the 50 k-point model with <3% error.

this approach allows a designer to reliably estimate the tail of the SNM PDF for a large memory with relatively few samples.

Thus far we have assumed that device mismatch occurs in transistors that start off as typical for the process. In addition to the inter-die V_T mismatch that we have described is an intra-die process variation that sets the process corner (e.g., fast nMOS, slow pMOS, etc.). Even for no mismatch, the process corner impacts the SNM. Fig. 18 shows the SNM PDF for a minimum sized 6T bitcell from a M-C simulation of global process corner in which nine process parameters are varied. Here again, the tail of the PDF is the limiting factor.

In a production framework, each die containing a given SRAM will have a global process corner that affects SNM as in Fig. 18. On top of this, mismatch in each cell will result from random doping variation. Assuming that any die within 3σ of the mean is usable, we found the global process corner that gives an SNM yield with the same probability as -3σ for both hold and read cases. Fig. 19 shows that the impact of mismatch at this 3σ process corner is essentially to shift the mean of the PDF by the offset caused by global variation. This means that the models we have presented remain valid for the case of combined global and local variation. Fig. 20 shows the semilog plot of the distributions to confirm this conclusion.

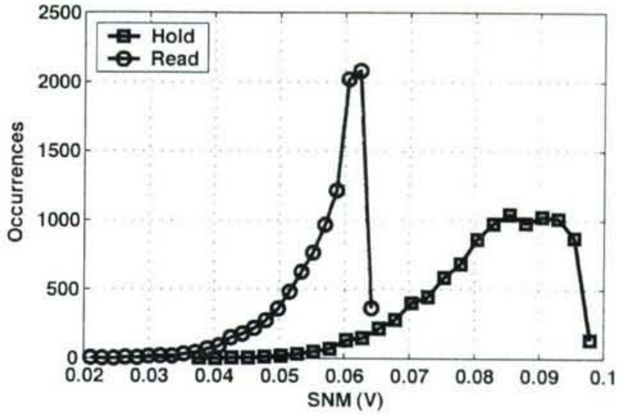


Fig. 18. Monte Carlo simulation showing global variation impact on SNM for a minimum sized bitcell.

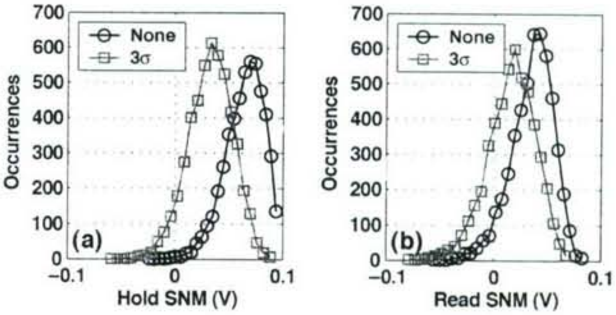


Fig. 19. SNM Monte Carlo simulations for local mismatch on top of global variation.

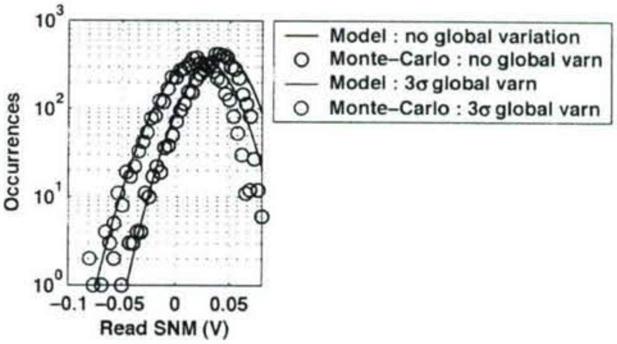


Fig. 20. SNM Monte Carlo simulations for local mismatch on top of global variation compared to the model.

III. Conclusion

Static noise margin is a critical metric for SRAM bitcell stability. This paper has explored the impact of different parameters on SNM for SRAM bitcells in sub-threshold. The dominant factor affecting sub-threshold circuits in general and SNM specifically is V_T mismatch due to random doping variation, and the critical region for examination is the tail of the SNM PDF. We have shown that first-order theoretical models for calculating SNM are accurate close to the nominal values of V_T, but they cannot accurately account for all of the mismatch cases.

We have shown that SNM high and SNM low are normally distributed with V_T mismatch and correlated. Despite their correlation, we have shown that treating them as *iid* leads to a PDF for SNM that gives an accurate model of the tail cases. This estimate is invaluable for avoiding long M-C simulations in the design of large SRAMs for sub-threshold operation.

REFERENCES

[1] A. Wang and A. Chandrakasan, "A 180 mV FFT processor using sub-threshold circuit techniques," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2004, pp. 292–293.

[2] M. Yamaoka, Y. Shinozaki, N. Maeda, Y. Shimazaki, K. Kato, S. Shimada, K. Yanagisawa, and K. Osadal, "A 300 MHz 25 μA/Mb leakage on-chip SRAM module featuring process-variation immunity and low-leakage-active mode for mobile-phone application processor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2004, pp. 494–495.

[3] N. Kim, K. Flautner, D. Blaauw, and T. Mudge, "Circuit and microarchitectural techniques for reducing cache leakage power," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, no. 2, pp. 167–184, Feb. 2004.

[4] A. Bhavnagarwala, S. Kosonocky, S. Kowalczyk, R. Joshi, Y. Chan, U. Srinivasan, and J. Wadhwa, "A transregional CMOS SRAM with single, logic VDD and dynamic power rails," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2004, pp. 292–293.

[5] E. Seevinck, F. List, and J. Lohstroh, "Static noise margin analysis of MOS SRAM cells," *IEEE J. Solid-State Circuits*, vol. SC-22, no. 5, pp. 748–754, Oct. 1987.

[6] B. Calhoun and A. Chandrakasan, "Analyzing static noise margin for sub-threshold SRAM in 65 nm CMOS," in *Proc. Eur. Solid-State Circuits Conf. (ESSCIRC)*, Sep. 2005, pp. 363–366.

[7] R. M. Swanson and J. D. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits," *IEEE J. Solid-State Circuits*, vol. SC-7, no. 2, pp. 146–153, Apr. 1972.

[8] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM leakage suppression by minimizing standby supply voltage," in *Int. Symp. Quality Electronic Design (ISQED) Dig. Tech. Papers*, 2004, pp. 55–60.

[9] A. Bhavnagarwala, X. Tang, and J. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE J. Solid-State Circuits*, vol. 36, no. 4, pp. 658–665, Apr. 2001.

[10] E. Vittoz, "Weak inversion for ultimate low-power logic," in *Low-Power Electronics Design*, C. Piguet, Ed. Boca Raton, FL: CRC Press, 2005, pp. 16-1–16-18.

[11] B. Cheng, S. Roy, and A. Asenov, "The impact of random doping effects on CMOS SRAM cell," in *Proc. Eur. Solid-State Circuits Conf. (ESSCIRC)*, 2004, pp. 219–222.

[12] R. Keyes, "The effect of randomness in the distribution of impurity atoms on FET threshold," *Appl. Phys. A: Mater. Sci. Process.*, vol. 8, pp. 251–259, 1975.

[13] S. Mukhopadhyay, H. Mahmoodi-Meimand, and K. Roy, "Modeling and estimation of failure probability due to parameter variations in nano-scale SRAMs for yield enhancement," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2004, pp. 64–67.

[14] M. Yamaoka, K. Osada, R. Tsuchiya, M. Horiuchi, S. Kimura, and T. Kawahara, "Low power SRAM menu for SOC application using yin-yang-feedback memory cell technology," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2004, pp. 288–291.

[15] K. Takeuchi, R. Koh, and T. Mogami, "A study of threshold voltage variation for ultra-small bulk and SOI CMOS," *IEEE Trans. Electron Devices*, vol. 48, no. 9, pp. 1995–2001, Sep. 2001.

Benton H. Calhoun (S'05–M'06) received the B.S. degree in electrical engineering with a concentration in computer science from the University of Virginia, Charlottesville, in 2000. He received the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, in 2002 and 2006, respectively.

In January 2006, he joined the faculty of the University of Virginia as an Assistant Professor in the Electrical and Computer Engineering Department. His research interests include leakage reduction, sensor networks, energy-efficient circuits, memory design, and sub-threshold operation.

Anantha P. Chandrakasan (M'95–SM'01–F'04) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from the University of California, Berkeley, in 1989, 1990, and 1994, respectively.

Since September 1994, he has been with the Massachusetts Institute of Technology, Cambridge, where he is currently the Joseph F. and Nancy P. Keithley Professor of Electrical Engineering. His research interests include low-power digital integrated circuit design, wireless microsensors, ultra-wideband radios, and emerging technologies. He is a coauthor of *Low Power Digital CMOS Design* (Kluwer, 1995) and *Digital Integrated Circuits* (Pearson Prentice-Hall, 2003, 2nd edition). He is also a co-editor of *Low Power CMOS Design* (IEEE Press, 1998), *Design of High-Performance Microprocessor Circuits* (IEEE Press, 2000), and *Leakage in Nanometer CMOS Technologies* (Springer, 2005).

Dr. Chandrakasan has received several awards, including the 1993 IEEE Communications Society's Best Tutorial Paper Award, the IEEE Electron Devices Society's 1997 Paul Rappaport Award for the Best Paper in an EDS publication during 1997, the 1999 Design Automation Conference Design Contest Award, and the 2004 DAC/ISSCC Student Design Contest Award. He has served as a technical program co-chair for the 1997 International Symposium on Low Power Electronics and Design (ISLPED), VLSI Design'98, and the 1998 IEEE Workshop on Signal Processing Systems. He was the Signal Processing Subcommittee Chair for ISSCC 1999–2001, the Program Vice-Chair for ISSCC 2002, the Program Chair for ISSCC 2003, and the Technology Directions Subcommittee Chair for ISSCC 2004–2006. He was an Associate Editor for the IEEE JOURNAL OF SOLID-STATE CIRCUITS from 1998 to 2001. He serves on the SSCS AdCom and is the meetings committee chair. He is the Technology Directions Chair for ISSCC 2007.

A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation

Benton Highsmith Calhoun, *Member, IEEE*, and Anantha P. Chandrakasan, *Fellow, IEEE*

***Abstract*—Low-voltage operation for memories is attractive because of lower leakage power and active energy, but the challenges of SRAM design tend to increase at lower voltage. This paper explores the limits of low-voltage operation for traditional six-transistor (6 T) SRAM and proposes an alternative bitcell that functions to much lower voltages. Measurements confirm that a 256-kb 65-nm SRAM test chip using the proposed bitcell operates into sub-threshold to below 400 mV. At this low voltage, the memory offers substantial power and energy savings at the cost of speed, making it well-suited to energy-constrained applications. The paper provides measured data and analysis on the limiting effects for voltage scaling for the test chip.**

***Index Terms*—Low-voltage memory, sub-threshold SRAM, voltage scaling.**

I. Introduction

SUBTHRESHOLD digital circuit design has emerged as a low-energy solution for applications with strict energy constraints. Analysis of sub-threshold designs has focused on logic circuits (e.g., [1]). SRAMs comprise a significant percentage of the total area and total power for many digital chips [2]. SRAM leakage can dominate total chip leakage, and switching highly capacitive bitlines and wordlines is costly in terms of energy. Lowering V_{DD} for SRAM saves leakage power and access energy. Also, for system integration, SRAM must become capable of operating at sub-threshold voltages that are compatible with sub-threshold combinational logic. Overcoming the difficulties of operating an SRAM in sub-threshold requires both circuit and architectural innovations. The benefits are significant, however, since low-energy SRAM is essential for enabling ultra-low-energy systems. This paper describes an SRAM capable of operating in the sub-threshold region.

Previous low-power memories show a trend of lower voltage operation. Exploiting dynamic voltage scaling (DVS) for SRAM is one motivation for designing a voltage-scalable memory. A 0.18-μm 32-kB four-way associative cache offers DVS compatability from 120 MHz, 1.7 mW at 0.65 V to 1.04 GHz, 530 mW at 2 V [3]. Although DVS can provide power reduction for active memories, most previous approaches apply voltage scaling primarily to idle blocks by lowering V_{DD} (e.g., [2], [4]–[6]), raising ground (e.g., [7]–[10]), or both (e.g., [11]). Implementations of SRAM using lower V_{DD} in standby are available [5] along with software policies to determine when to enter the lower leakage mode [2]. Voltage scaling for SRAM promises to continue, leading to sub-threshold storage modes and even sub-threshold operation for SRAMs operating in tandem with sub-threshold logic.

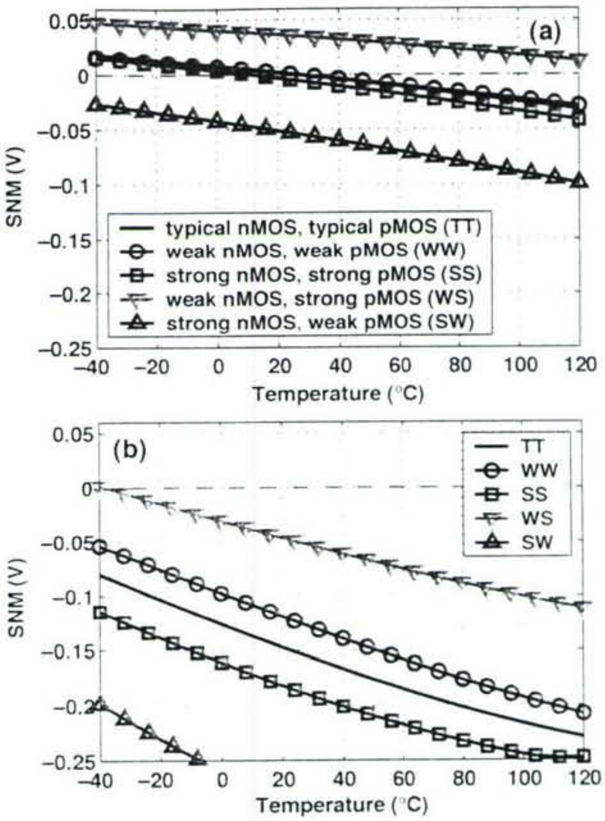


Fig. 1. SNM for write access versus temperature and process corner (TT, WW, SS, WS, and SW) at $V_{DD} = 0.3$ V (a) and $V_{DD} = 0.6$ V (b). Negative SNM indicates successful write.

One issue for deeply voltage scaled SRAM is soft error rate (SER). Soft errors occur when an alpha particle or cosmic ray strikes a memory node and causes data loss. Since bitcell storage capacitance decreases with scaling and voltage scaling further reduces the stored charge, SER is a concern for sub-threshold memory. Fortunately, there are methods for handling soft errors. Studies of soft errors have shown that multi-cell errors from a single strike only occur in two to three adjacent cells along a wordline [12]. Thus, physically interspersing bits from different words can prevent multi-errors from occurring in a single word [12]. Coupling this with error correcting codes can dramatically reduce SER [8].

Manuscript received May 23, 2006; revised August 1, 2006. This work was supported by DARPA and by Texas Instruments Incorporated. Chip fabrication was provided by Texas Instruments.

B. H. Calhoun is with the Electrical and Computer Engineering Department, University of Virginia, Charlottesville, VA 22904-4743 USA (e-mail: bcalhoun@virginia.edu).

A. P. Chandrakasan is with the Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Digital Object Identifier 10.1109/JSSC.2006.891726

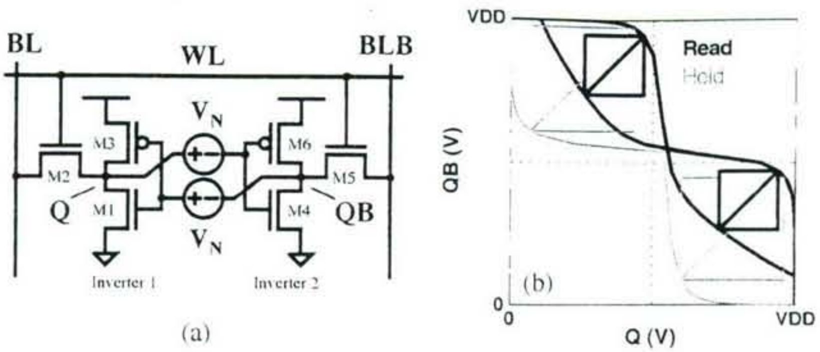


Fig. 2. Schematic for 6 T bitcell showing voltage noise sources for finding SNM [15] (a), and example Hold and Read SNM butterfly plots (b).

Clearly, previous efforts have explored many options for voltage scaling. However, none have yet pushed voltage scaling into the sub-threshold region during active operation.

II. Six-Transistor SRAM Bitcell in Sub-threshold

Predictions in [13] suggest that process variations will limit standard 90-nm SRAMs to around 0.7 V operation due to degraded Read Static Noise Margin (SNM) and reduced write margin. Small transistors combine with random and systematic process variations to cause a large spread in Read SNM that leads to destructive read errors for bits at the tail of the distribution. Standard write operation depends on a ratio of currents, and process variations make this ratio difficult to maintain as V_{DD} decreases, leading to write errors. These practical problems limit traditional six-transistor (6 T) bitcells and architectures to higher V_{DD}, above-threshold operation. Reports in the literature of 65-nm SRAMs confirm this voltage barrier. A 65-nm SRAM built in a dynamic-double-gate SOI (D2G-SOI) process functions to 0.7 V and is predicted to fail below 1.0 V for bulk CMOS [14]. A bulk CMOS 65-nm SRAM reports a minimum operating voltage of 0.7 V [9]. Our results confirm that SNM degradation and inability to write are the two primary obstacles to sub-threshold SRAM functionality, where they are exacerbated by the exponential impact of V_T variations.

A. *Write Operation*

Proper write operation depends on sizing the access nMOS to win the ratioed fight with the pMOS inside the bitcell to write a "0". For a successful write, the bitcell becomes monostable, forcing the internal voltages to the correct values. If the cell retains bistability then the write does not occur, and the SNM is positive on the cell's butterfly plot. Thus, a negative SNM indicates a successful write (monostability in the cell). For above-V_T operation, stronger nMOS devices (due to mobility) and relatively low dependence of current on V_T make device sizing successful at maintaining the proper ratio of currents for writing the cell. For sub-threshold, the ratio of currents in p/nMOS depends exponentially on V_T. Since process designers generally focus on strong-inversion operation, the sub-threshold pMOS and nMOS current can be imbalanced for typical transistors. Even if the pMOS and nMOS currents are well-balanced at the typical nMOS, typical pMOS (TT) corner, process variation can still create a relative difference in p/nMOS current of an order of magnitude or more. Furthermore, local variations in V_T from cell to cell can aggravate this problem. For sub-V_T, sizing alone is not a strong knob for fixing this problem because only unreasonable sizing ratios could account for the wide ranges of possible current that arise due to V_T mismatch.

In the 65-nm process for which we are designing, iso-size pMOS devices are stronger in sub-V_T than nMOS by roughly an order of magnitude, which makes write functionality more challenging. Fig. 1 shows the write margin (neg. SNM means successful write) of a 6 T bitcell versus temperature and process corner. At $V_{DD} = 300$ mV in Fig. 1(a), the writing fails for large regions of process corner and temperature. The general trend showing an improvement of write operation (i.e., more negative margin) at higher temperature occurs because the pMOS transistors weaken relative to nMOS as temperature rises. As V_{DD} increases, the write margin improves. Fig. 1(b) shows the write margin at 0.6 V. This voltage is above V_T, so the pMOS has weakened relative to the nMOS because the mobility dominates the differences in V_T. Even at 0.6 V, the write margin is barely negative for the worst-case corner, and this plot does not account for local V_T variation. For these reasons, $V_{DD} = 0.6$ V is the best case voltage for which we can expect traditional write operations to work for a sub-threshold memory in this 65-nm process.

B. *Read Operation: Static Noise Margin*

Fig. 2 shows a conceptual setup for modeling SNM [15]. Noise sources having value V_N are introduced at each of the internal nodes in the bitcell. As V_N increases, the stability of the cell reduces. Once V_N exceeds the SNM, then the cell loses its bistability and its data. Cell stability during active operation represents a more significant limitation to SRAM operation than during hold. At the onset of a read access, the wordline is "1" and the bitlines are precharged to "1". The internal node of the bitcell that represents a zero gets pulled upward through the access transistor due to the voltage dividing effect across the access transistor (M_2, M_5) and drive transistor (M_1, M_4), which degrades the Read SNM. Fig. 2 shows example butterfly curves during hold and read that illustrate the degradation in SNM during read.

Process variation makes matters worse by shifting the voltage transfer characteristics (VTCs) of the cell inverters and creating a distribution of SNM for both hold and read. A study of the impact of variations on SNM in sub-V_T appears in [16]. Fig. 3

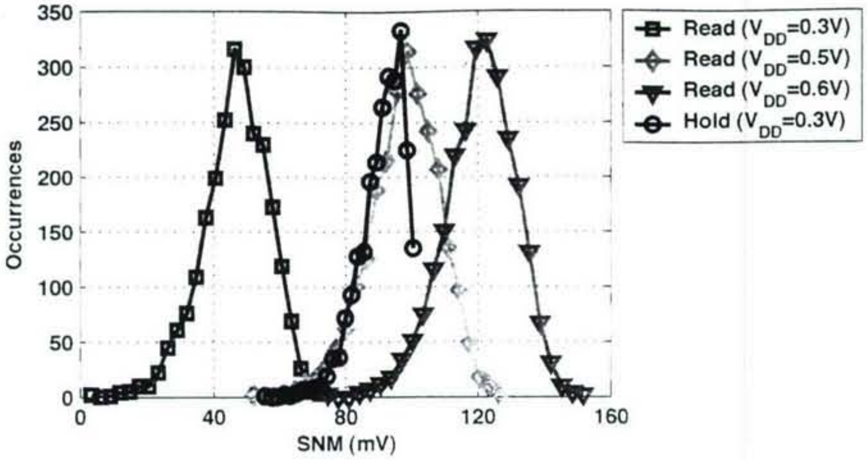


Fig. 3. Distribution of Hold SNM at 300 mV compared with Read SNM distributions at different voltages. Read SNM at 500 mV has the same mean, but it has a larger standard deviation.

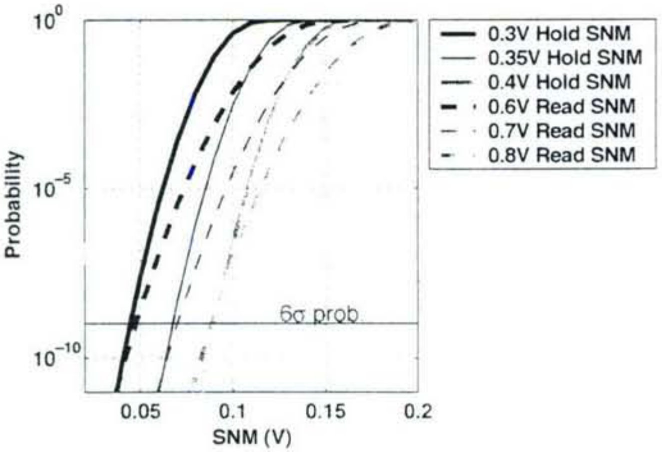


Fig. 4. CDFs of SNM distributions showing that avoiding the Read SNM allows a reduction in V_{DD} by ~0.5 for the same 6σ stability.

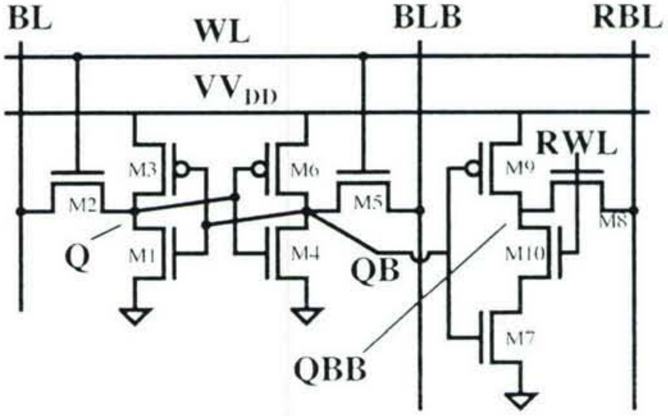


Fig. 5. Schematic of the 10 T sub-threshold bitcell.

shows the distribution of the Read and Hold SNMs for a 6 T bitcell at a 300-mV supply voltage. The mean Read SNM is only slightly above half of the mean Hold SNM, and the deviation of the Read SNM is larger than for the Hold SNM. For a multiple megabit memory, numerous cells will have Read SNM less than zero based on this statistical analysis. From this figure, the mean of the Read SNM at 500 mV roughly equals the mean of the Hold SNM at 300 mV. However, it is unclear from this plot how the Hold SNM and Read SNM compare at the worst-case tails. Fig. 4 shows the cumulative distribution function (CDFs) derived from the distributions. For 6σ probability, the Hold SNM for a given V_{DD} roughly equals the Read SNM for twice that V_{DD} in the range of interest. This means that a memory that avoids the Read SNM problem can operate at roughly half of the V_{DD} of a 6 T memory with the same 6σ bitcell stability.

III. A Sub-threshold Bitcell Design

Previously published works have scaled SRAM V_{DD} into the sub-threshold region during idle, but no SRAM actually operates in this region. The 0.18-μm memory in [1] provides one exception, operating into deep sub-V_T at 180 mV. However, the memory resembles a register file (latch with tristate driver for writing and muxed outputs) and has an equivalent bitcell size of 18 T. We can use this previous implementation [1] as an endpoint in the range of bitcell options that spans from the 6 T bitcell (inoperable below 600–700 mV in 65 nm) to the 18 T bitcell, which will function robustly in sub-threshold since it looks and functions more like combinational logic. In between these two options are many possible bitcell designs that address the obstacles to sub-threshold operation by increasing the number of transistors relative to the 6 T cell. The bitcell that this section describes [17] was selected from among many others because it represents the best trade-off of functionality and area; it is the smallest bitcell from those examined that provides robust sub-threshold functionality.

Fig. 5 shows the schematic of the 10 T sub-threshold bitcell. Transistors M_1 through M_6 are identical to a 6 T bitcell except that the source of M_3 and M_6 tie to a virtual supply voltage rail, VV_{DD}. Write access to the bitcell occurs through the write access transistors, M_2 and M_5, from the write bitlines, BL and BLB. Transistors M_7 through M_{10} implement a buffer used for reading. Read access is single-ended and occurs on a separate bitline, RBL, which is precharged to V_{DD} prior to read access. The wordline for read also is distinct from the write wordline. One key advantage to separating the read and write wordlines and bitlines is that a memory using this bitcell can have distinct

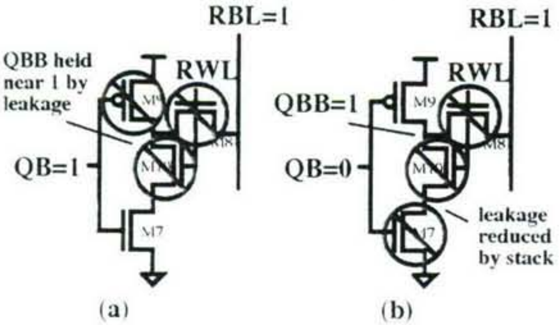


Fig. 6. Schematic of read buffer from 10 T bitcell for both data values. In both cases, leakage is reduced to the bitline and through the inverter relative to the case where M_{10} is excluded.

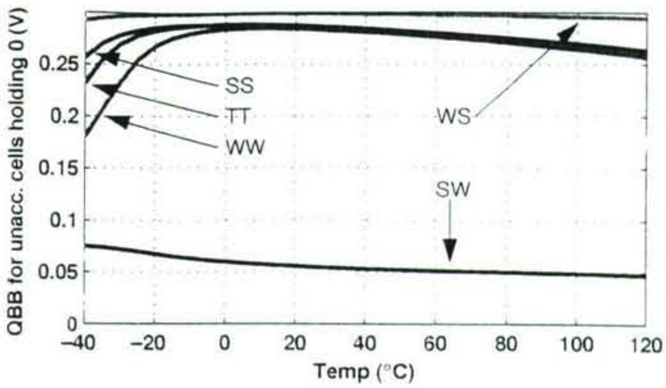


Fig. 7. Simulation of voltage at node QBB in unaccessed 10 T bitcells versus temperature and process corner. Strong pMOS leakage holds QBB near V_{DD} except at the SW corner. Even at SW, QBB is higher than it is for the 6 T cell, lowering bitline leakage.

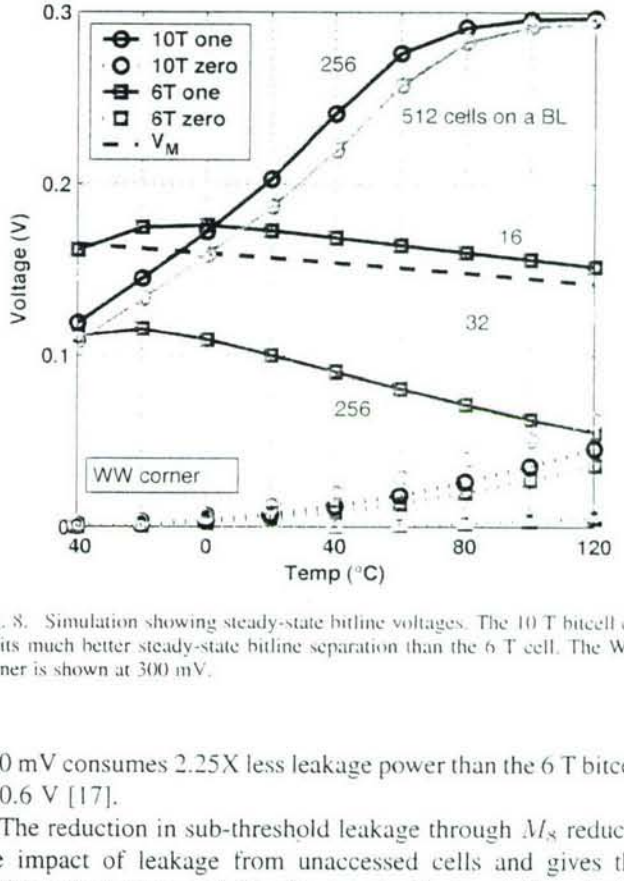


Fig. 8. Simulation showing steady-state bitline voltages. The 10 T bitcell exhibits much better steady-state bitline separation than the 6 T cell. The WW corner is shown at 300 mV.

read and write ports. Since a 6 T bitcell does not have this feature, the 10 T bitcell is in some ways more fairly compared to an 8 T dual-port bitcell (6 T bitcell with two pairs of access transistors and bitlines).

A. *Enabling Sub-threshold Read*

The 10 T bitcell in Fig. 5 uses transistors M_7–M_{10} to remove the problem of Read SNM by buffering the stored data during a read access. As described previously, eliminating the Read SNM problem allows this bitcell to operate at half of the V_{DD} of a 6 T cell while retaining the same 6σ stability. A different approach for eliminating the Read SNM in [18] uses a 7 T cell to prevent the higher voltage at the internal node from propagating to the other back-to-back inverter by holding its data dynamically during read accesses. This approach will not work in sub-V_T because the dynamic data is susceptible to leaking away during the long access times.

It is interesting to note that a 9 T bitcell, identical to the bitcell in Fig. 5 but without M_{10}, would eliminate the Read SNM problem while using less area than the 10 T cell. However, M_{10} is valuable to the bitcell because it reduces leakage current and allows more bitcells to share a bitline. Fig. 6 shows the read buffer from the 10 T bitcell for $Q = 0$ (a) and $Q = 1$ (b). When $Q = 0$ and $QB = 1$ Fig. 6(a), M_{10} adds an *off* device in series with the leakage path through M_8 and the path through M_9, decreasing the leakage through those transistors. Furthermore, since the pMOS in this 65-nm technology generally has higher leakage than the nMOS, the leakage in M_9 holds node QBB near V_{DD} (see Fig. 7), further limiting the leakage through M_8 by making its V_{GS} negative. Even if QBB floats above 0 by only a small amount, the negative V_{GS} in M_8 reduces bitline leakage exponentially. When $Q = 1$ and $QB = 0$ Fig. 6(b), M_{10} reduces leakage through M_7 by the stack effect (note that the stack of devices will also slow down a read access by decreasing read current). Since node QBB is held solidly at V_{DD}, M_8 has $V_{DS} = 0$, so bitline leakage is negligible. In both cases, M_{10} reduces the leakage relative to the 9 T (and 6 T) case. The 10 T only has 16% more leakage than a 6 T cell at the same V_{DD} (9 T has 50% more). This overhead in leakage current is more than compensated by decreasing V_{DD} by several hundred millivolts relative to the 6 T bitcell. In simulation, the 10 T bitcell at 300 mV consumes 2.25X less leakage power than the 6 T bitcell at 0.6 V [17].

The reduction in sub-threshold leakage through M_8 reduces the impact of leakage from unaccessed cells and gives the additional advantage of allowing more cells on a bitline during read. Leakage from the bitline into the unaccessed bitcells causes undesirable voltage drop that slows differential sensing and that makes single-ended read values difficult to distinguish. Fig. 8 shows the impact of bitline leakage on steady-state voltages (note that the bitline initially is precharged to V_{DD}) while reading a "1" (solid lines) or "0" (dotted lines) at 300 mV. For the same number of cells on a BL, the 10 T bitcell (circles) shows larger bitline separation than the 6 T (or 9 T) bitcells (squares). This figure suggests that "sensing" with an inverter (whose switching threshold, V_M, is shown) should work well from 0 °C to 100 °C even with 256 cells on a bitline for the 10 T cell. In contrast, the 6 T cell (or 9 T bitcell) would allow at most 16 bitcells on a bitline. The bitline that should be "1" stays very

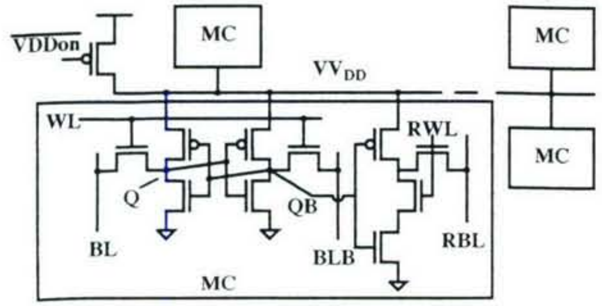


Fig. 9. Schematic of write architecture for a single row using a floating power supply ($VV_{\rm DD}$). The row is "folded" in layout so that its cells share n-wells, and the entire row is written at once.

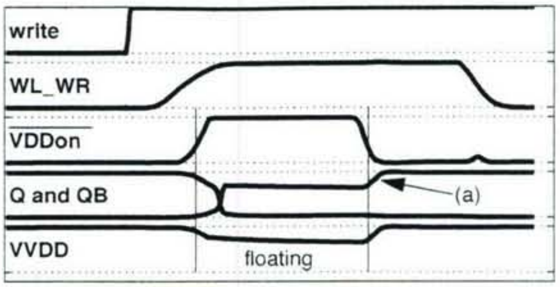


Fig. 10. Timing diagram for write operation. When $\overline{V_{DDon}}$ goes low while WL_{WR} remains asserted, the cell's feedback restores full voltage levels for the new values of Q and QB (point (a)).

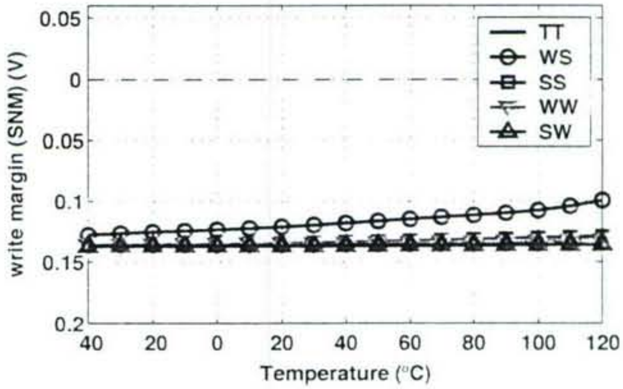


Fig. 11. Write margin (write SNM) versus temperature at 0.3 V for 10 T bitcell with floating $VV_{\rm DD}$ supply. Negative margin for all corners, signifying successful write operation.

close to $V_{\rm DD}$ at high temperatures and then begins to droop at lower temperatures. This occurs because M_{10} inside the unaccessed 10 T bitcells is so successful at reducing sub-V_T current through the access transistors that the sub-V_T current actually drops below the sum of gate currents (which is fairly constant with temperature) into unaccessed cells. If gate leakage was lower (e.g., high-K dielectrics), then sub-threshold leakage into the unaccessed cells is reduced sufficiently such that the bitline will stay very close to $V_{\rm DD}$. One advantage of more cells on a BL is a reduction in peripheral circuits that offsets some of the area overhead of larger bitcells. For example, an 8 T bitcell (40% larger than 6 T) that allows 256 cells per bitline rather than 16 (same improvement as our cell) actually resulted in a 6% smaller overall array area [19].

B. Enabling Sub-threshold Write

In this 65-nm technology, a 6 T bitcell cannot write in the traditional fashion below around 0.6 V because the nMOS access transistor cannot reliably win the ratioed fight against the pMOS to write a "0". The technique of weakening the cross-coupled inverters by gating their supply voltage (e.g., [6]) or ground node (e.g., [20]), applied by previous works primarily to improve speed, can dramatically improve write margin. Fig. 9 shows the schematic for a single row using this approach. A single power-supply-gating header switch connects node $VV_{\rm DD}$ to the true power rail. When the bitcell holds its data or during read accesses, $\overline{V_{DDon}} = 0$ so that $VV_{\rm DD} = V_{\rm DD}$. During a write access, the virtual rail floats. For the implementation on the test chip, a conceptual row folds as shown in the figure so that its bitcells can share n-wells, and the entire row is written at once.

Fig. 10 shows the timing associated with a write access using this scheme. First, the write signal goes high to indicate that a write access will occur, and the bitlines are driven with the new data. Next, the decoders drive a global wordline (not shown) which causes the proper local write wordline (WL_{WR}) to go high. Triggered by the local wordline, the $\overline{V_{DDon}}$ signal goes high, allowing node $VV_{\rm DD}$ to float. As the write access transistors discharge the virtual rail, its voltage droops, and Q and QB change to their new values. The logical "1" inside the cell tracks the drooping voltage until $\overline{V_{DDon}}$ goes low again while the *local wordline remains high*, and the virtual rail reconnects to $V_{\rm DD}$. The feedback inside the bitcell then holds the Q and QB nodes at their correct logical values and amplifies the "1" to full $V_{\rm DD}$ (point (a) in Fig. 10). The plot in Fig. 11 shows the write margin for the virtual $V_{\rm DD}$ approach across temperature and process corner at $V_{\rm DD} = 300$ mV. The margin remains negative across all of these ranges, indicating a successful write.

IV. 65-nm Sub-threshold SRAM Test Chip

A. Test Chip Architecture

A 256-kb 65-nm bulk CMOS test chip uses the 10 T bitcell and the architecture shown in Fig. 12. The memory has eight 32-kb blocks with 256 rows and 128 columns each. A single 128-bit DIO bus serves all eight blocks. In this initial instantiation of the sub-threshold memory, only one read or write can occur per cycle, however the 10 T bitcell would allow a read and write access to the same block in one cycle. Such a dual-port instantiation of the memory would require a second DIO bus and additional peripheral logic. A combined global wordline and block select signal assert a local wordline that triggers either WL_{RD} or WL_{WR}. For a write access, $M_P\langle r\rangle$ for the accessed row turns off. The write drivers consist simply of inverters with transmission gates, which turn off when the memory is not writing to minimize leakage on the write bitlines (BL and BLB). The power supply to the WL drivers is routed separately to allow a boosted WL voltage. This technique improves the access speed and increases the robustness to local variations. The read bitline (RBL) is precharged prior to read access, and its steady-state value is "sensed" using a simple inverter, $I_{RD}\langle c\rangle$. Column and row redundancy is a ubiquitous technique in commercial memories used to improve yield. For

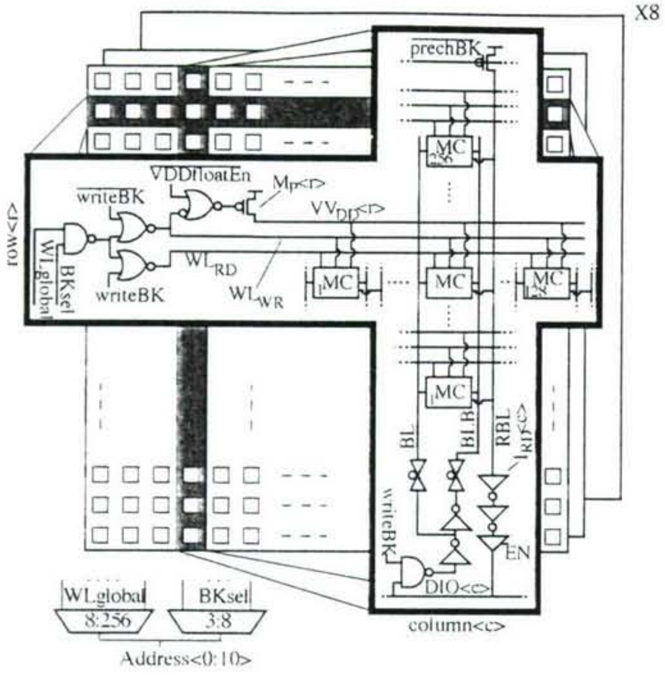


Fig. 12. Architecture diagram of the 256-kb memory on the test chip using 10 T sub-threshold bitcells.

our analysis of the SRAM, we assume the availability of one redundant row and column per block.

The primary goals for this test chip were to test the functionality of the 10 T bitcell in sub-threshold and to explore the limitations of the design. For this reason, all of the peripherals use static CMOS logic for simplicity and for functional robustness in sub-threshold. The large block size was intentionally aggressive to expose limitations in the bitcell and architecture. Integrating 256 bitcells on the bitline (as opposed to 16 for 6 T) pushes the envelope for functionality. The 10 T bitcell layout added 66% area overhead relative to our reference 6 T design, but the overall area penalty will be less due to more bitcells on a bitline, as described in Section III-A. Each row is folded such that a pair of 64-bit physical rows sharing n-wells and a VV_{DD} rail makes up one conceptual 128-bit row (c.f. Fig. 9). This folding increases the length of bitlines by roughly 2X and decreases the length of wordlines by roughly (1/2)X. Notice that this is not fundamentally necessary for the write approach to work. The n-wells of two separate rows can be shared and the VV_{DD} for each row routed separately. The 10 T architecture does not change the number of WLs or BLs or the number of devices per line relative to the 6 T case, except that it has one fewer read BL. The capacitance of the metal lines themselves will increase somewhat due to the larger bitcell area. Fig. 13 shows a layout shot and die photograph of the test chip (1.89 mm by 1.12 mm, pin-limited).

B. Measurements

Measurements of the SRAM test chip confirm that it is functional over a range of voltages from 1.2 V down well into the sub-threshold region. With the assumption of one redundant row and column per block, read operation works without error to 320 mV and write operation works without error to 380 mV at 27 °C. We continued to push the supply voltage to even lower values to examine the limits of the implementation. At the low supply voltage of 300 mV, the memory continues to function, but it exhibits bit errors in ~1% of its bits that result from sensitivities in the architecture to local device variation, as described later.

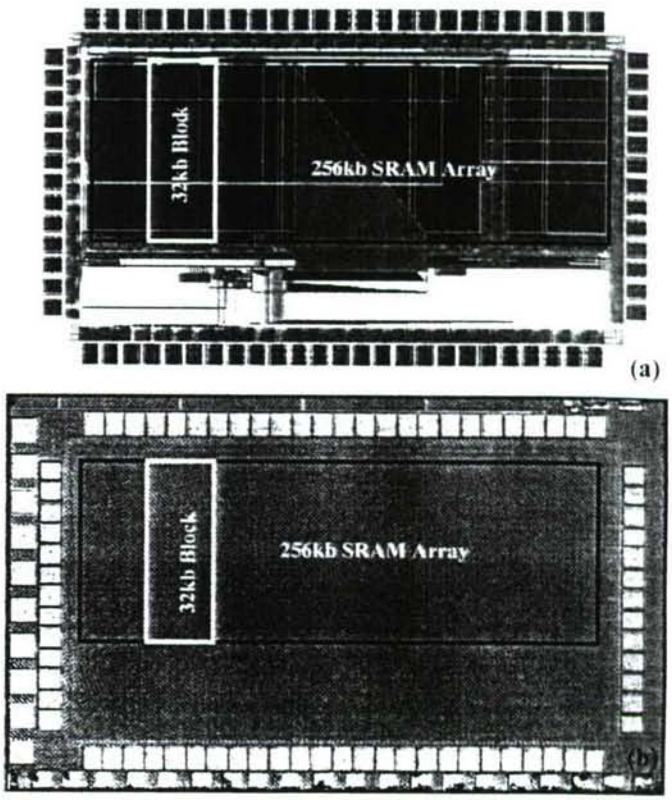


Fig. 13. (a) Annotated layout and (b) die photograph of the 256-kb sub-threshold SRAM in 65 nm. Die size is 1.89 mm by 1.12 mm.

The test chip successfully demonstrates a functional sub-threshold memory that overcomes the problems it was designed to face. First, the bitcell removes the Read SNM problem. Measurements have confirmed that the memory experiences zero destructive read errors at 300 mV. Simulations show that a 6 T memory would experience a high rate of destructive read errors at 300 mV due to degraded Read SNM. Second, whereas a 6 T memory would fail to write below about 600 mV, this memory writes correctly at 350 mV at 85 °C. Third, a 6 T memory would experience problems reading with only 16 bitcells on a bitline. Measurements show that the 10 T memory reads correctly even with 256 bitcells on the bitline down to 320 mV. Finally, the memory shows good Hold SNM performance. The first bits observed to fail to hold their data occur at $V_{DD} < 230$ mV, as seen in the distribution shown in Fig. 14.

Fig. 15 shows the measured leakage power of the test chip at two different temperatures and the expected savings from V_{DD} scaling. At 27 °C, the 10 T memory saves 2.5X and 3.8X in leakage power by scaling from 0.6 V to 0.4 V and 0.3 V, respectively and over 60X when V_{DD} scales from 1.2 V to 0.3 V. V_{DD} scaling also gives the expected savings in active energy per read

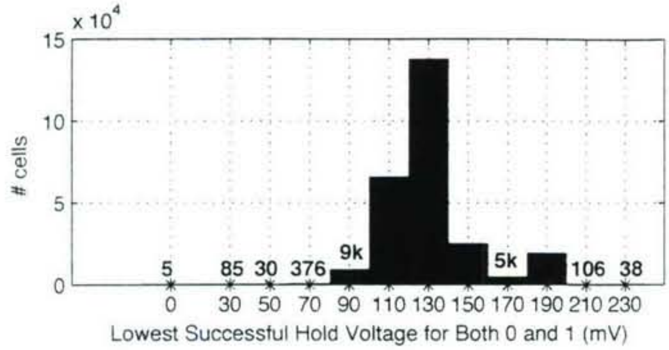


Fig. 14. Measured distribution of minimum voltage at which bitcells hold both "0" and "1".

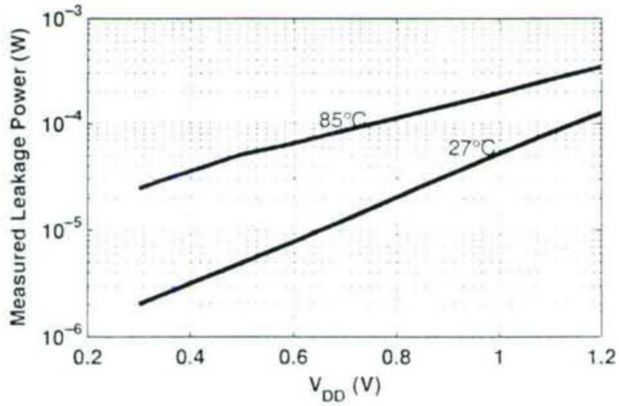


Fig. 15. Measured leakage power from the memory test chip.

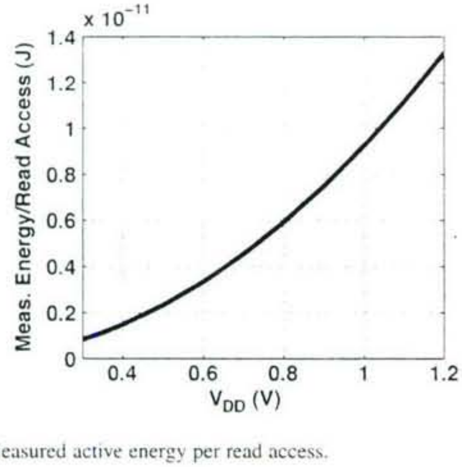


Fig. 16. Measured active energy per read access.

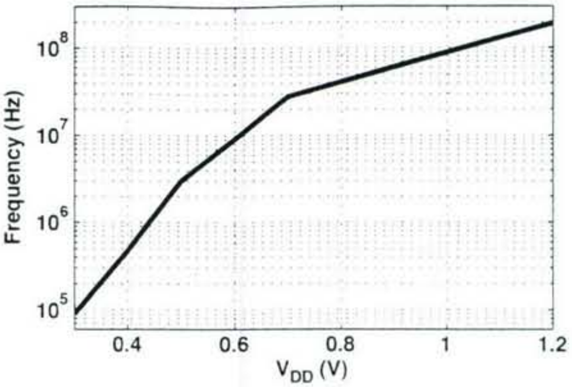


Fig. 17. Measured frequency of operation versus $V_{\rm DD}$.

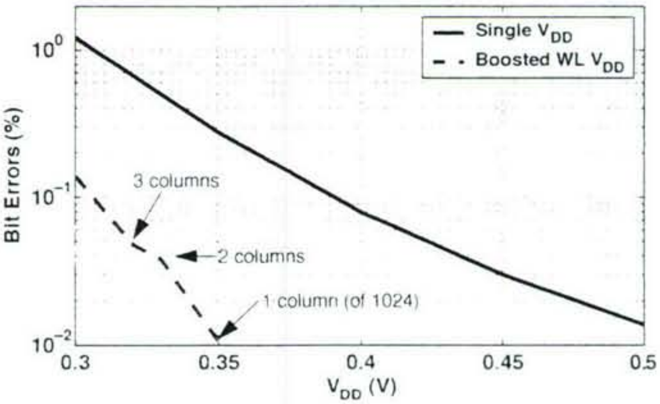


Fig. 18. Measured percentage of bit errors for read versus $V_{\rm DD}$. Boosting the WL voltage dramatically reduces these errors.

access, as shown in Fig. 16. Fig. 17 shows the measured frequency of operation versus $V_{\rm DD}$ (the 1.2 V speed of 200 MHz is a simulation result, because the testing board did not support high-speed testing). The maximum measured operating speed at 400 mV is 475 kHz.

Pushing $V_{\rm DD}$ even lower exposes the limitations for both read and write operations as a small fraction of bits begins to fail. These failures occur for the same set of bits in a repeatable fashion. For certain bits at low voltage, a read access shows that the bitcell holds a "0" when in fact it holds a "1". This error is non-destructive, which we confirm by raising the supply voltage and re-reading the cell and invariably reading the correct value. Also, these bit errors tend to gather along a small number of specific columns. The fact that this error exists in a small fraction of cases indicates that it results from local device variation, and we can isolate the problematic transistors. The fact that the bits exhibiting problems cluster along specific columns indicates that variation in the sensing inverter, I_{RD} in Fig. 12, has shifted its switching threshold, V_M, towards $V_{\rm DD}$. Now, specific bitcells along this column that have read access transistors weakened by local variation cannot hold the read bitline above V_M of inverter I_{RD}. Several experiments confirm that this is the mechanism for read failures. First, we can independently lower the supply voltage of the sense inverters, $I_{RD}\langle c\rangle$. This lowers the V_M for the inverters, and the measured bit error rate decreases. Secondly, we can increase temperature, which provides the expected improvement in discerning a "1" (c.f. Fig. 8). Finally, we can increase the voltage of the wordline drivers, which pulls the transistors that weakened by local variation back toward the mean and rapidly decreases the read bit errors. Fig. 18 shows the measured percentage of bit errors during read access versus $V_{\rm DD}$. The error rate without the wordline boosted by 100 mV also is shown. In sub-threshold, the extra gate voltage on the read access transistors provides over 10X (due to the sub-threshold slope) of extra current drive. As with above-threshold memories, the extra current speeds operation. It also makes the design more robust to mismatch.

By aggressively choosing a block having 256 rows on a single bitline, we pushed the limits of read operation and exposed the limits to scaling read accesses that result from local variation. Also, using a simple inverter for sensing makes it harder to read a "1" correctly. As the bitline separation plots have shown, V_M of the sensing inverter lies too close to the logical "1" value at some corners and temperatures. Boosting the wordline voltage offers one simple change that dramatically reduces the error rate and allows this memory to read without error at 320 mV. A better solution to improve the read reliability and robustness to local device variation is to replace the inverter with a new sensing scheme, for which many relevant sense amps are available in the literature.

The limit to V_{DD} scaling for write manifests when write accesses fail for specific bitcells. Write functionality was tested using a high voltage write, a low-voltage write of the opposite value, and finally a high voltage read. This test isolates the bits for which sub-threshold write fails. These errors aggregate in bits along specific rows. As with the read errors, local device variation is the culprit, and the predominance of row-wise errors suggests that the failure mechanism involves the row peripherals. Referring back to Fig. 12, write limitations first appear along specific rows whose pull-up device, M_P, is strengthened by local variation. Thus, when M_P turns off during a write access, its larger leakage pulls VV_{DD} up closer to V_{DD}. Some of the bitcells can still switch under these conditions, but VV_{DD} reaches a steady-state voltage that is high enough to prevent some bitcells from overpowering the pMOS to write a '0' into the memory. In these bitcells, local mismatch has made the internal pMOS relatively stronger than the access transistor to the point that the write driver cannot flip the cell at the steady-state VV_{DD} voltage. Measurements confirm that this is the case. First, the lowest functional supply voltage decreases at higher temperature. Since the leakage through M_P gets relatively weaker compared with the nMOS access transistors, this confirms the mechanism for failure. More importantly, the write errors decrease when the supply voltage to the wordline increases. The higher wordline voltage increases V_{GS} for the write access transistors and makes them more capable of producing voltage droop on VV_{DD}. Fig. 19 shows the percentage of bit errors measured during write both with and without 100 mV of wordline boosting. With boosting, the memory can write without error at 380 mV at 27 °C and 350 mV at 85 °C.

As with the limitations on read, simple changes to the peripheral circuits can push the lowest operational V_{DD} even lower. Specifically, the leakage through M_P can be reduced using one of several well-known methods (e.g., stacking, RBB, etc.) A better solution to the write issue that maintains the same basic architecture and approach is to induce a specific voltage drop on VV_{DD} intentionally. In the extreme, replacing M_P with an inverter will drive VV_{DD} all the way to 0 V. Then, as long as the write wordline remains asserted, the bitcells will develop the correct internal data when VV_{DD} goes back high regardless of local variations. A disadvantage of this extreme case is the energy penalty associated with discharging and re-powering the VV_{DD} rail and all of the bitcells in the row. An alternative is to use a circuit (e.g., diode connected FET) to force VV_{DD} to some intermediate value that is low enough to ensure write but that uses less energy.

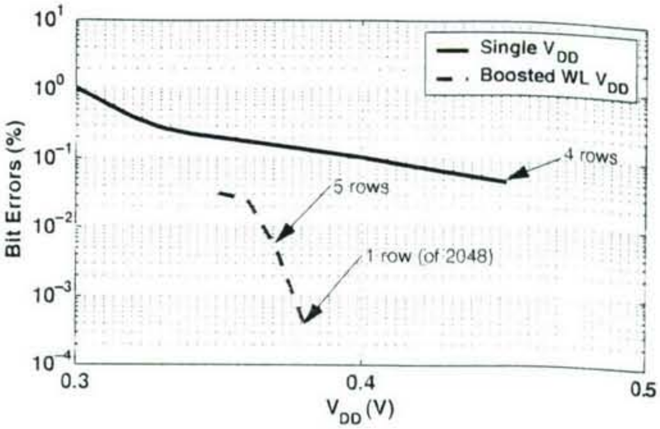


Fig. 19. Measured percentage of bit errors for write versus V_{DD}. Again, boosting the WL voltage dramatically reduces these errors.

V. Summary and Conclusions

Sub-threshold SRAM provides the dual advantages of minimizing total memory energy consumption and of providing compatability with minimum-energy sub-threshold logic. Traditional 6 T SRAM cannot function in sub-threshold because it fails to write and because the Read SNM degrades badly. Furthermore, bitline leakage in 6 T SRAMs limits the number of bitcells on a bitline to 16. Measurements of a 256-kb 65-nm bulk CMOS test chip show that our 10 T bitcell fundamentally solves the Read SNM problem, overcomes the write problem, and relaxes the bitline integration limitation to allow sub-threshold operation. With one redundant row and column per block and a boosted wordline, the memory functions without error to below 380 mV. At 400 mV, it consumes 3.28 μW and works up to 475 kHz. Although aggressive design exposes the limitations of the architecture in terms of its robustness to local device variation, the bit errors result primarily from problems in the peripheral circuits. Simple proposed changes to the periphery promise to push the limits of SRAM operation to even lower V_{DD}.

References

[1] A. Wang and A. Chandrakasan, "A 180 mV FFT processor using sub-threshold circuit techniques," in *IEEE ISSCC Dig. Tech. Papers*, 2004, pp. 292–293.

[2] N. Kim, K. Flautner, D. Blaauw, and T. Mudge, "Circuit and microarchitectural techniques for reducing cache leakage power," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, no. 2, pp. 167–184, Feb. 2004.

[3] K. Osada, J. L. Shin, M. Khan, Y. Liou, K. Wang, K. Shoji, K. Kuroda, S. Ikeda, and K. Ishibashi, "Universal-Vdd 0.65–2.0-V 32-kB cache using a voltage-adapted timing-generation scheme and a lithographically symmetrical cell," *IEEE J. Solid-State Circuits*, vol. 36, no. 11, pp. 1738–1744, Nov. 2001.

[4] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM leakage suppression by minimizing standby supply voltage," in *Proc. ISQED*, 2004, pp. 55–60.

[5] K. Kanda, T. Miyazaki, M. K. Sik, H. Kawaguchi, and T. Sakurai, "Two orders of magnitude leakage power reduction of low-voltage SRAM's by row-by-row dynamic V_{DD} control (RRDV) scheme," in *Proc. IEEE Int. ASIC/SOC Conf.*, Sep. 2002, pp. 381–385.

[6] A. Bhavnagarwala, S. Kosonocky, S. Kowalczyk, R. Joshi, Y. Chan, U. Srinivasan, and J. Wadhwa, "A transregional CMOS SRAM with single, logic V_{DD} and dynamic power rails," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2004, pp. 292–293.

[7] H. Yamauchi, T. Iwata, H. Akamatsu, and A. Matsuzawa, "A 0.8 V/100 MHz/sub-5 mW-operated mega-bit SRAM cell architecture with charge-recycle offset-source driving (OSD) scheme," in *Symp. VLSI Circuits Dig. Tech. Papers*, 1996, pp. 126–127.

[8] K. Osada, Y. Saitoh, E. Ibe, and K. Ishibashi, "16.7-fA/cell tunnel-leakage-suppressed 16-Mb SRAM for handling cosmic-ray-induced multierrors," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1952–1957, Nov. 2003.

[9] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Yang, B. Zheng, and M. Bohr, "A SRAM design on 65 nm CMOS technology with integrated leakage scheme," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2004, pp. 294–295.

[10] A. Agarwal, H. Li, and K. Roy, "A single-V_t low-leakage gated-ground cache for deep submicron," *IEEE J. Solid-State Circuits*, vol. 38, no. 2, pp. 319–328, Feb. 2003.

[11] T. Enomoto, Y. Oka, and H. Shikano, "A self-controllable voltage level (SVL) circuit and its low-power high-speed CMOS circuit applications," *IEEE J. Solid-State Circuits*, vol. 38, no. 7, pp. 1220–1226, Jul. 2003.

[12] K. Osada, K. Yamaguchi, Y. Saitoh, and T. Kawahara, "SRAM immunity to cosmic-ray-induced multierrors based on analysis of an induced parasitic bipolar effect," *IEEE J. Solid-State Circuits*, vol. 39, no. 5, pp. 827–833, May 2004.

[13] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. Shimazaki, K. Nii, S. Shimada, K. Yanagisawa, and T. Kawahara, "Low-power embedded SRAM modules with expanded margins for writing," in *IEEE ISSCC Dig. Tech. Papers*, 2005, pp. 480–481.

[14] M. Yamaoka, K. Osada, R. Tsuchiya, M. Horiuchi, S. Kimura, and T. Kawahara, "Low power SRAM menu for SOC application using Yin-Yang-feedback memory cell technology," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2004, pp. 288–291.

[15] E. Seevinck, F. List, and J. Lohstroh, "Static noise margin analysis of MOS SRAM cells," *IEEE J. Solid-State Circuits*, vol. SC-22, no. 5, pp. 748–754, Oct. 1987.

[16] B. Calhoun and A. Chandrakasan, "Analyzing static noise margin for sub-threshold SRAM in 65 nm CMOS," in *Proc. ESSCIRC*, 2005, pp. 363–366.

[17] B. Calhoun and A. Chandrakasan, "A 256 kb sub-threshold SRAM in 65 nm CMOS," in *IEEE ISSCC Dig. Tech. Papers*, 2006, pp. 628–629.

[18] K. Takeda, Y. Hagihara, Y. Aimoto, M. Nomura, Y. Nakazawa, T. Ishii, and H. Kobatake, "A read-static-noise-margin-free SRAM cell for low-Vdd and high-speed applications," in *IEEE ISSCC Dig. Tech. Papers*, 2005, pp. 478–479.

[19] A. Alvandpour, D. Somasekhar, R. Krishnamurthy, V. De, S. Borkar, and C. Svensson, "Bitline leakage equalization for sub-100 nm caches," in *Proc. ESSCIRC*, 2003, pp. 401–404.

[20] K. Itoh, A. Fridi, A. Bellaouar, and M. Elmasry, "A deep sub-V, single power-supply SRAM cell with multi- V_T, boosted storage node and dynamic load," in *Symp. VLSI Circuits Dig. Tech. Papers*, 1996, pp. 132–133.

Benton Highsmith Calhoun (S'05–M'06) received the B.S. degree in electrical engineering with a concentration in computer science from the University of Virginia, Charlottesville, VA, in 2000, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, in 2002 and 2006, respectively.

In January 2006, he joined the faculty at the University of Virginia as an Assistant Professor in the Electrical and Computer Engineering Department. His research interests include low-power digital circuit design, sub-threshold digital circuits, SRAM design for end-of-the-roadmap silicon, variation tolerant circuit design methodologies, and low-energy electronics for medical applications. He is a coauthor of *Sub-threshold Design for Ultra Low-Power Systems* (Springer, 2006).

Dr. Calhoun serves on the Technical Program Committee for the International Symposium on Low Power Electronics and Design (ISLPED).

Anantha P. Chandrakasan (M'95–SM'01–F'04) received the B.S, M.S., and Ph.D. degrees in electrical engineering and computer sciences from the University of California, Berkeley, in 1989, 1990, and 1994, respectively.

Since September 1994, he has been with the Massachusetts Institute of Technology, Cambridge, where he is currently the Joseph F. and Nancy P. Keithley Professor of Electrical Engineering. His research interests include low-power digital integrated circuit design, wireless microsensors, ultra-wideband radios, and emerging technologies. He is a coauthor of *Low Power Digital CMOS Design* (Kluwer, 1995) and *Digital Integrated Circuits* (Pearson Prentice-Hall, 2003, 2nd edition). He is also a co-editor of *Low Power CMOS Design* (IEEE Press, 1998), *Design of High-Performance Microprocessor Circuits* (IEEE Press, 2000), *Leakage in Nanometer CMOS Technologies* (Springer, 2005), *Sub-threshold Design for Ultra Low-Power Systems* (Springer, 2006).

Dr. Chandrakasan has received several awards including the 1993 IEEE Communications Society's Best Tutorial Paper Award, the IEEE Electron Devices Society's 1997 Paul Rappaport Award for the Best Paper in an EDS publication during 1997, the 1999 Design Automation Conference Design Contest Award, and the 2004 DAC/ISSCC Student Design Contest Award. He has served as a technical program co-chair for the 1997 International Symposium on Low Power Electronics and Design (ISLPED), VLSI Design'98, and the 1998 IEEE Workshop on Signal Processing Systems. He was the Signal Processing Subcommittee Chair for ISSCC 1999–2001, the Program Vice-Chair for ISSCC 2002, the Program Chair for ISSCC 2003, and the Technology Directions Subcommittee Chair for ISSCC 2004–2006. He was an Associate Editor for the IEEE JOURNAL OF SOLID-STATE CIRCUITS from 1998 to 2001. He serves on the SSCS AdCom and is the meetings committee chair. He is the Technology Directions Chair for ISSCC 2007.

500-MS/s 5-bit ADC in 65-nm CMOS With Split Capacitor Array DAC

Brian P. Ginsburg, *Student Member, IEEE*, and Anantha P. Chandrakasan, *Fellow, IEEE*

***Abstract*—A 500-MS/s 5-bit ADC for UWB applications has been fabricated in a 65-nm CMOS technology using no analog-specific processing options. The time-interleaved successive approximation register (SAR) architecture has been chosen due to its simplicity versus flash and its amenability to scaled technologies versus pipelined, which relies on operational amplifiers. Six time-interleaved channels are used, sharing a single clock operating at the composite sampling rate. Each channel has a split capacitor array that reduces switching energy, increases speed, and has similar INL and decreased DNL, as compared to a conventional binary-weighted array. A variable delay line adjusts the instant of latch strobing to reduce preamplifier currents. The ADC achieves Nyquist performance, with an SNDR of 27.8 and 26.1 dB for 3.3 and 239 MHz inputs, respectively. The total active area is 0.9 mm^2, and the ADC consumes 6 mW from a 1.2-V supply.**

***Index Terms*—ADC, analog-to-digital conversion, deep-submicron CMOS, successive approximation register, ultra-wideband radio.**

I. Introduction

ULTRA-WIDEBAND (UWB) radio is an emerging technology for very-high-data-rate, short distance wireless communications. Both OFDM [1] and pulse-based [2] solutions are being developed to achieve data rates in excess of 480 Mb/s. UWB receivers require high-speed but low-resolution analog-to-digital converters (ADCs), in the range of 4–5 bits [3]–[5]. The ADC in this work is targeted for specifications (5 bit, 500 MS/s) compatible with a custom pulse-based UWB transceiver [6], [7], where 100 Mb/s communication is achieved using BPSK-modulated 500-MHz-wide Gaussian pulses transmitted in one of 14 bands between 3.1–10.6 GHz.

The flash topology, along with its interpolating and folding variants, has been the conventional choice for high-speed, low-resolution ADCs [8]–[12]. While flash can maintain the highest throughput, it requires an exponential growth in the number of comparisons with the resolution. The ensuing complexity motivates the use of other architectures.

Pipelined ADCs are used for high-speed, medium-resolution applications [13], [14]. They can provide one conversion per clock period throughput and only a linear scaling in complexity with resolution; however, they rely on operational amplifiers at the heart of the multiplying digital-to-analog converter (MDAC) in each pipelined stage. Because it must be closed loop stable, this amplifier typically uses one or two high gain stages. Unfortunately, in deep-submicron CMOS, the achievable gain per stage is limited because short-channel effects lower $g_m r_o$ for a single transistor, and reduced voltage supplies restrict circuit techniques such as cascoding. Thus, there are significant challenges for continued scaling of pipelined ADCs.

Very recently, for the high-speed, low resolution converters necessary for UWB, the time-interleaved successive approximation register (SAR) architecture has re-emerged[1] as a low-power alternative to flash and pipelined ADCs [17]. At the required speeds, their major limitation is digital power; a SAR converter includes digital feedback in the critical path. A full custom logic controller with dynamic registers can reduce digital power significantly, but it still remains a dominant source of power consumption in a 0.18-μm CMOS implementation [18]. Another approach uses dynamic registers with asynchronous operation to reduce clock power, and combined with a non-binary successive approximation algorithm, has led to a very energy efficient design in 0.13-μm CMOS [19]. Fortunately, technology scaling improves the digital power and speed without many of the issues plaguing pipelined converters. The only active analog component in a SAR ADC, the comparator, still requires large gain and bandwidth, but because it does not have to be linear, this gain can be achieved through cascaded stages and positive feedback.

This paper presents a 500-MS/s 5-bit ADC fabricated in a 65-nm CMOS technology [20]. At the maximum sampling rate, the ADC consumes 6 mW from a 1.2-V supply. This low power consumption is achieved through proper architecture selection, a new capacitor array, and careful timing allocation between the digital and analog circuits. The ADC has six time-interleaved SAR channels synchronized to a common clock. The split capacitor array reduces switching energy, is robust to digital delay mismatches for overall improved settling time, and has a reduction in peak static differential nonlinearity (DNL). In the comparator, a variable delay line adjusts the instant of strobing for the regenerative latches, minimizing idle time during each bit-cycle without sacrificing bit error rate (BER) performance.

II. ADC Architecture

A SAR ADC requires one period for sampling and b periods to resolve the b digital output bits. To make the internal SAR clock synchronous to the overall sampling clock, six time-interleaved channels are used, as shown in Fig. 1. Thus, only a single 500 MHz clock is required in the prototype, easing clock generation and distribution. The channels synchronize by passing

Manuscript received August 25, 2006; revised December 19, 2006. This work was supported by the Defense Advanced Research Projects Agency (DARPA) and a National Defense Science and Engineering Graduate (NDSEG) Fellowship.

The authors are with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: bginzz@mit.edu).

Digital Object Identifier 10.1109/JSSC.2007.892169

[1] Time-interleaved SAR was used as early as 1980 as a low area alternative to the flash ADC [15], and, more recently, for reduced comparator power in a medium resolution application [16].

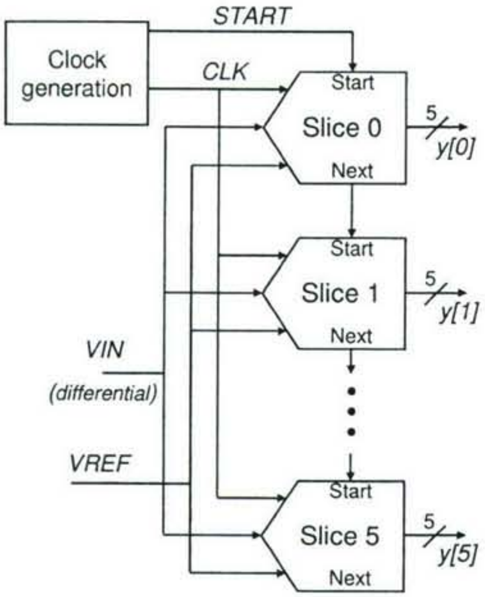


Fig. 1. Top-level block diagram of the 6-way time-interleaved ADC.

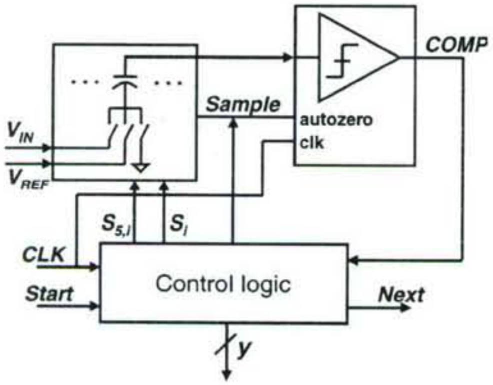


Fig. 2. Block diagram of the channel, which has a capacitive DAC, comparator, and digital logic.

a token to cue their start of sampling, and all critical sampling edges are aligned to the same shared clock [18]. Timing skew between channels is thus limited to routing variations to the channels and the delay mismatch through a single register in each channel; both of these error sources can be kept sufficiently small such that digital timing correction (a complex, power hungry process [21]) is not necessary.

The channel, shown in Fig. 2, consists of a capacitive digital-to-analog converter (DAC), a comparator, and control logic (itself called the SAR). The control logic switches the DAC using a binary search algorithm to minimize the error between the digital output and the analog input. The split capacitor array and comparator, the two analog blocks, are discussed in Section III, followed by some of the considerations used in designing circuits for 65-nm CMOS.

III. Circuit Design

A. *Split Capacitor Array*

The DAC serves two purposes in a SAR converter: it samples the input charge, and it generates an error voltage between the input and current digital estimate. The conventional DAC choice is a binary-weighted capacitor array [22], as shown in Fig. 3, which is insensitive to stray capacitance. As shown in [23], however, the conventional capacitor array uses charge inefficiently during a conversion. To demonstrate this, a conversion of a 2-bit capacitor array is presented here. During the first bit decision after sampling, the MSB capacitor is connected to V_{REF} with the remaining capacitors connected to ground (left circuit in Fig. 4). The output of the capacitor array, V_X, is

$$V_X = -V_{\mathrm{IN}} + \frac{1}{2}V_{\mathrm{REF}} \tag{1}$$

where V_{IN} is the input voltage sampled on the capacitor array and V_{REF} is the reference voltage. During the second bit-cycle, the SAR does one of two transitions. If $V_X < 0$, an "up" transition is performed, where C_1 is switched from ground up to V_{REF}, drawing

$$E_{\mathrm{up}} = \frac{C_0 V_{\mathrm{REF}}^2}{4} \tag{2}$$

from the reference voltage supply. Inversely, if $V_X > 0$, a "down" transition is performed (Fig. 4); C_1 and C_2 switch places. If they switch at the same time, the energy required is

$$E_{\mathrm{down,conv}} = \frac{5}{4}C_0 V_{\mathrm{REF}}^2. \tag{3}$$

It takes 5 times more energy to lower V_X than to raise it; this occurs because all of the charge initially on C_2 is discharged to ground, and all the charge that ends up on C_1 must be delivered from the reference voltage supply.

Ref. [23] analyzes three alternatives to the conventional capacitor array and switching procedure. Of these alternatives, this work implements the split capacitor array because it has both the lowest switching energy and does not require an extra clock phase that would limit high speed operation. A b-bit split capacitor array is shown in Fig. 5; the MSB capacitor of the conventional array has been split into an identical copy (MSB subarray) of the rest of the array (main subarray). These arrays are placed in parallel (common top plate), not to be confused with the series connected capacitor arrays used in the sub-DAC approach.[2] The total capacitance of the split capacitor array is $2^b C_0$, identical to the conventional case, and the area requirements are unchanged.

The split capacitor switching algorithm is presented in Fig. 6. Here, the two-bit example from above is repeated for the split capacitor array to demonstrate the switching method and energy savings. During the first bit-cycle (left side of Fig. 7), the MSB subarray, $C_{2,1}$ and $C_{2,0}$, is connected to V_{REF}, and the main subarray is connected to ground. Since $C_2 = C_{2,1} + C_{2,0}$, (1) also represents the output of the split array. In the case of an "up" transition, the array transitions in the same method as above, with C_1 switching to V_{REF}, consuming the same energy calculated in (2). In the "down" transition (Fig. 7), half of the MSB subarray, $C_{2,1}$ is lowered to ground, leaving both C_1 and $C_{2,0}$ unchanged. By only switching one capacitor the energy

[2] Historically, the combination of capacitive main- and sub-DACs had been called a "split array" [15], but this has not become common usage, and we have co-opted the term for the new structure.

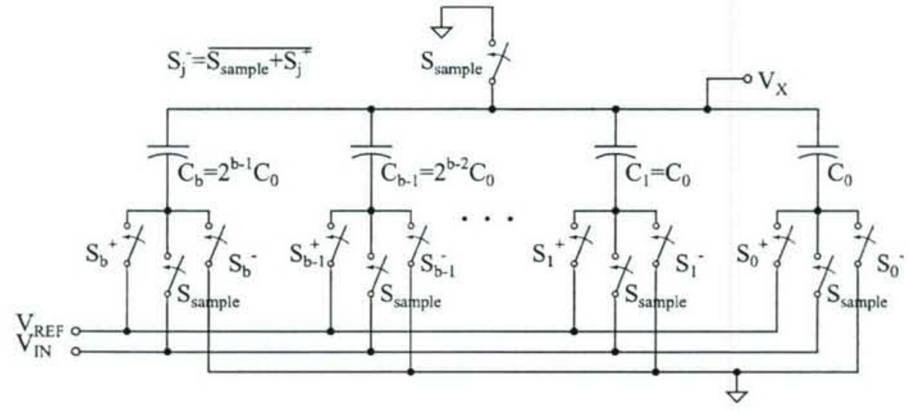


Fig. 3. Conventional b-bit binary weighted capacitor array.

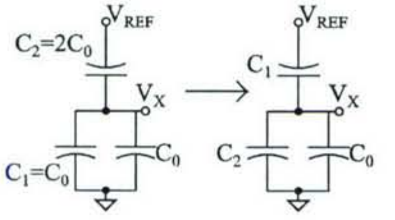


Fig. 4. "Down" transition of the conventional capacitor array.

consumed is

$$E_{\text{down,split}} = \frac{C_0 V_{\text{REF}}^2}{4} \tag{4}$$

identical to the "up" transition.

The overall energy savings of the split capacitor array is input voltage (or output digital code) dependent. Where the relative frequency of "down" transitions is greater, the savings for the split capacitor array is enhanced, as seen in Fig. 8. Assuming a full swing sinusoidal input distribution, the split capacitor array is expected to have 37% lower switching energy than the conventional array.

For this high-speed implementation, an additional advantage of considerable significance is related to the array's settling time. During a "down" transition, two capacitors are required to switch for the conventional capacitor array; any mismatch, whether random or deterministic, in the digital logic driving these switches can cause the capacitor array to initially transition in the wrong direction, potentially exacerbating an overdrive condition for the preamplifiers. Only one capacitor in the split capacitor array transitions during any bit-cycle, providing inherent immunity to the skew of the switch signals. Simulation results comparing the settling times of the two arrays is shown in Fig. 9. For the simulation, the total width of the switches is identical for the split and conventional arrays. The split capacitor array settles up to 10% faster, which is used to reduce the bias currents in the preamplifiers by a similar amount.

1) Linearity Performance: To compare the theoretical static linearity of the binary-weighted and split DACs, each of the capacitors is modeled as the sum of the nominal capacitance value and some error term:

$$\begin{aligned} C_n &= 2^{n-1}C_0 + \delta_n \\ C_{b,n} &= 2^{n-1}C_0 + \delta_{b,n}. \end{aligned} \tag{5}$$

Initially, consider only the case where all the errors are in the unit capacitors, whose values are independent identically-distributed (i.i.d.) Gaussian random variables; later in this section, other non-idealities will be considered. Then the error terms δ_n and $\delta_{b,n}$ have zero mean, are independent, and have variance

$$E\left[\delta_n^2\right] = E\left[\delta_{b,n}^2\right] = 2^{n-1}\sigma_0^2 \tag{6}$$

where σ_0 is the standard deviation of the unit capacitor.

The linearity of a SAR ADC is limited by the accuracy of the DAC outputs, which are calculated here for the case of no initial charge on the array ($V_{\text{IN}} = 0$). For a given DAC digital input $y = \sum_{n=1}^{b} S_n 2^{n-1}$, with S_n equals 0 or 1 represents the ADC decision for bit n, the analog output for the conventional binary-weighted array is

$$V_{X,conv}(y) = \frac{\sum\limits_{n=1}^{b}\left(2^{n-1}C_0 + \delta_n\right)S_n}{2^b C_0 + \Delta C} V_{\text{REF}}. \tag{7}$$

The second term in the denominator $\Delta C = \sum_{n=0}^{b} \delta_n$ will be neglected for this discussion. This will make the analysis simpler but will prevent a complete closed form solution for the integral nonlinearity (INL). Subtracting the nominal value yields the error term

$$V_{\text{err}}(y) \approx \frac{\sum\limits_{n=1}^{b} \delta_n S_n}{2^b C_0} V_{\text{REF}} \tag{8}$$

with variance

$$\begin{aligned} E\left[V_{\text{err}}^2(y)\right] &= \frac{\sum\limits_{n=1}^{b} 2^{n-1}\sigma_0^2 S_n}{2^{2b}C_0^2} V_{\text{REF}}^2 \\ &= \frac{y}{2^{2b}} \frac{\sigma_0^2}{C_0^2} V_{\text{REF}}^2. \end{aligned} \tag{9}$$

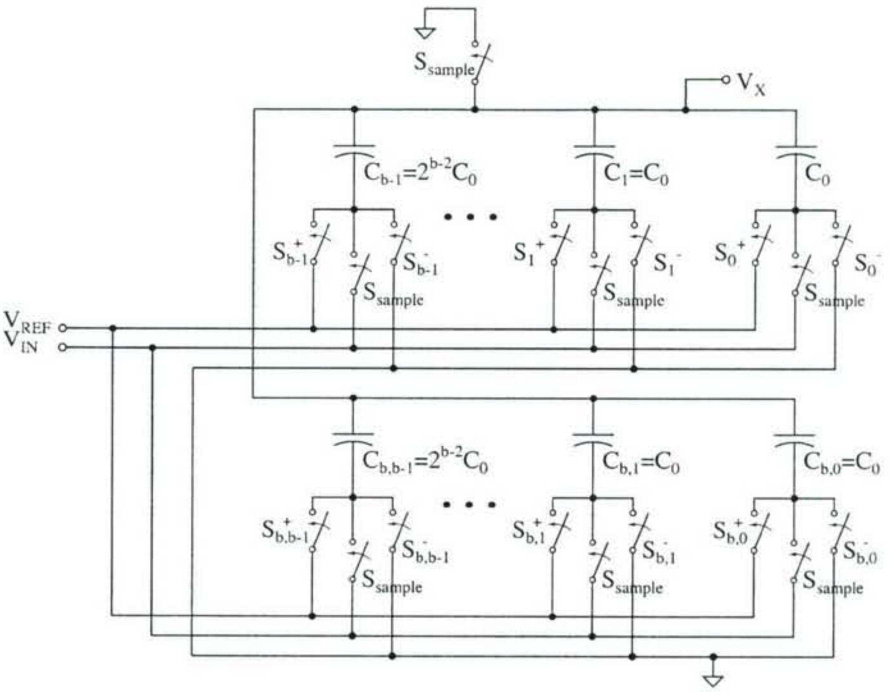


Fig. 5. The b-bit split capacitor array, with the main subarray on top and the MSB subarray below.

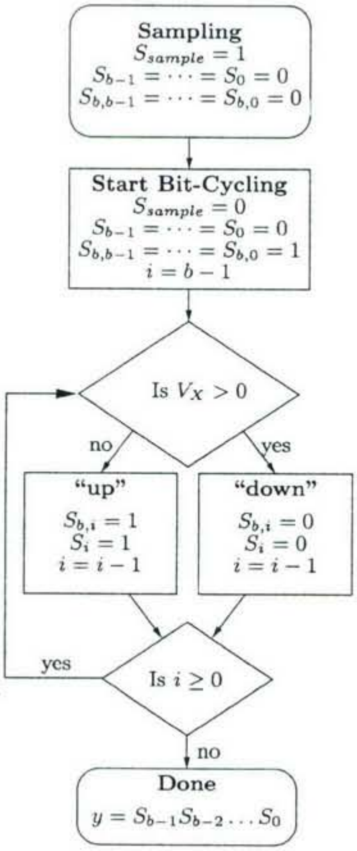


Fig. 6. Switching procedure for split capacitor array. i represents the bit currently being decided.

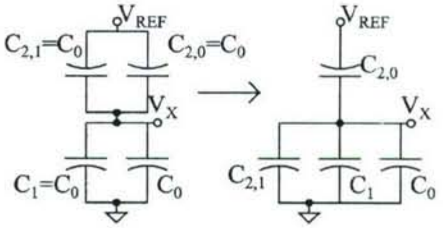


Fig. 7. "Down" transition of the split capacitor array. The "up" transition entails switching C_1 to V_{REF}.

This voltage error is simply the sum of the errors from y unit capacitors connected to V_{REF}. Because the errors in the unit capacitors are assumed to be i.i.d., it does not matter which unit capacitors are connected to V_{REF} but only the total number. Thus, (9) holds for the case of the split capacitor array as well. This error is also directly related to the INL of the ADC, and thus there should be no difference between the maximum INLs of the two arrays.

The DNL of the capacitive DAC is, neglecting gain errors, the difference between the voltage errors at two consecutive DAC outputs, as in

$$\mathrm{DNL}(y) \approx \Delta V_{\mathrm{err}}(y) = V_{\mathrm{err}}(y) - V_{\mathrm{err}}(y-1). \tag{10}$$

The worst case DNL for the binary weighted capacitor array is expected to occur at the step below the MSB transition, where its variance is

$$\begin{aligned} E\left[\Delta V_{\mathrm{err}}^2\left(2^{b-1}\right)\right] &= E\left[\left(\frac{\delta_b - \sum\limits_{n=1}^{b-1}\delta_n}{2^b C_0} V_{\mathrm{REF}}\right)^2\right] \\ &\approx \frac{\sigma_0^2}{2^b C_0^2} V_{\mathrm{REF}}^2. \end{aligned} \tag{11}$$

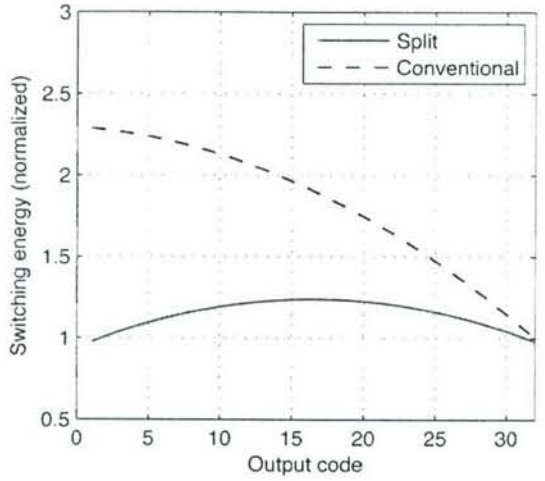


Fig. 8. Normalized switching energies of the conventional and split capacitor arrays versus output code. The number of "down" transitions is greater on the left side of the plot.

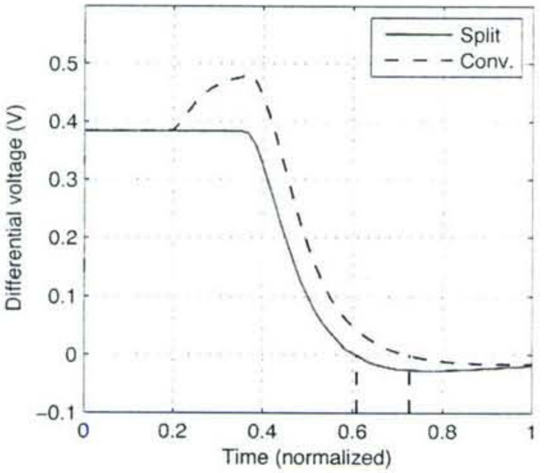


Fig. 9. Simulation of the settling time of the split and conventional capacitor arrays under the presence of digital timing skew.

For the split capacitor array, the worst case DNL also occurs at the step below the MSB transition, but its value is

$$\begin{aligned}\Delta V_{\text{err}}\left(2^{b-1}\right) &= \frac{\sum_{n=0}^{b-1}\delta_{b,n} - \left(\sum_{n=0}^{b-2}\delta_{b,n} + \sum_{n=1}^{b-2}\delta_n\right)}{2^b C_0} V_{\text{REF}} \\ &= \frac{\delta_{b,b-1} - \sum_{n=1}^{b-2}\delta_n}{2^b C_0} V_{\text{REF}}. \qquad (12)\end{aligned}$$

This error has a variance of

$$E\left[\Delta V_{\text{err}}^2\left(2^{b-1}\right)\right] \approx \frac{1}{2}\frac{\sigma_0^2}{2^b C_0^2} V_{\text{REF}}^2. \qquad (13)$$

Comparing (11) and (13) shows that the standard deviation of the worst case DNL is $\sqrt{2}$ lower for the split capacitor array. Conceptually, this occurs because the errors at $y = 2^{b-1}$ and $y = 2^{b-1} - 1$ are partially correlated for the split capacitor array, causing the cancellation of $\delta_{b,0}, \ldots, \delta_{b,b-2}$ in (12). This can be also be seen in the energy example above. In Fig. 4, the errors of the top capacitors are completely uncorrelated for the two bit decisions; however, in Fig. 7, the error of $C_{2,0}$ contributes equally to both bit decisions.

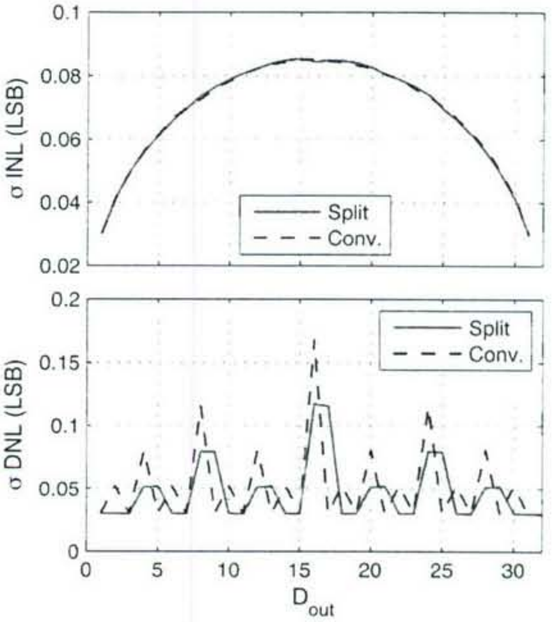


Fig. 10. Behavioral simulation comparing the linearity of the split and conventional capacitor arrays. 10 000 Monte Carlo runs were performed, with i.i.d. Gaussian errors in the unit capacitors ($\sigma_0/C_0 = 3\%$). The standard deviation of the INL and DNL are plotted.

A behavioral simulation of the SAR ADC, with both the binary weighted and split capacitor arrays, was performed. The values of the unit capacitors are taken to be Gaussian random variables with standard deviation of 3% ($\sigma_0/C_0 = 0.03$), and the ADC is otherwise ideal. Fig. 10 shows the results of 10 000 Monte Carlo runs, where the standard deviation of the INL and DNL are plotted versus output code at the 5-bit level. As expected, the conventional and split arrays have identical INL characteristics, and the split capacitor array has $\sqrt{2}$ better DNL. This improvement in DNL is similar to that conferred at the MSB transition from using 1-bit of unary decoding in a segmented DAC [24].

The above discussion assumes that the errors in the unit capacitors are due to an i.i.d. random process. In practice, care must be taken during layout to ensure absence of systematic nonidealities. The unit capacitors are arranged in a common centroid configuration to eliminate the effect of first order gradients. Fringing effects at the edge of the array are reduced by using 32 dummy capacitors around the 32 active unit capacitors. The largest capacitors in the main subarray and MSB subarray are distributed so as to have equal numbers of edges next to the dummy capacitors to further reduce fringing errors. The split capacitor array does have twice as many bottom plate signals that must be routed within the array. Coupling from these routes to the top plate routing can cause linearity errors and was avoided by routing the top and bottom plate signals distant from each other, which was sufficient at 5-bit resolution. For higher resolutions, electrostatic shielding may be necessary where the bottom

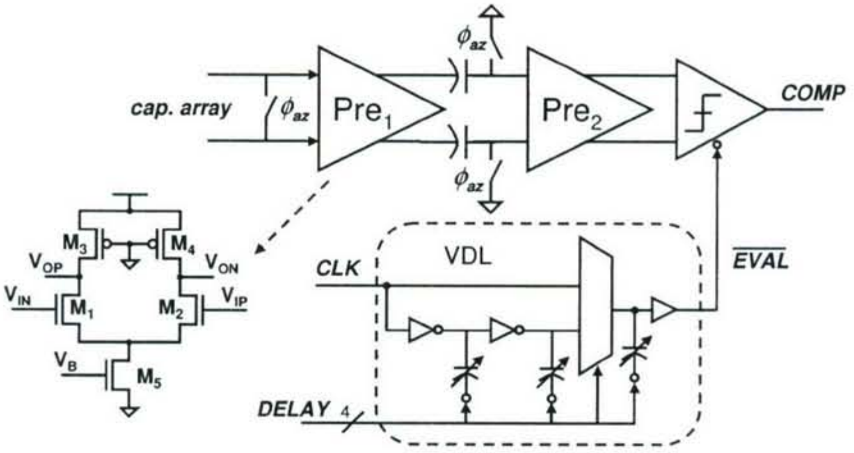


Fig. 11. Comparator schematic showing preamplifier chain, latch, and VDL inserted in series with the latch strobe signal.

plate routing is separated from the capacitors by grounded metal [25]. Shielding can also improve immunity to noise coupling from the substrate.

B. Comparator With Adjustable Strobing

The comparator, shown in Fig. 11, has a regenerative latch preceded by two stages of autozeroed preamplifiers, used to reduce the input referred offset of the latch to below one quarter of the LSB voltage. The preamplifiers are linear amplifiers with an input NFET differential pair M_1-M_2 and resistive loads, formed by PFETs M_3-M_4 operating in the linear region. The gain per stage is selected to be 3–4 for ease of integration at both low voltages and with very short channel devices. The offset of the first preamplifier is cancelled using output offset storage. The sizing of the preamplifiers, autozeroing capacitors, and latch follows the offset/matching-limited optimization procedure described in [26].

During bit-cycling, the clock period is divided into one phase for the settling of the DAC and preamplifiers and one phase for regeneration of the latch. The latch typically resolves, even for small inputs, in much less than the 1ns that is allocated assuming an even division of the period. The ADC sits idle after the latch settles until the start of the next bit-cycle. Self-timed bit-cycling uses this idle time to start the next bit-cycle early [18], [27]. This approach relaxes the preamplifier settling time requirement for all but the first bit-cycle (determining the MSB), as it has no prior bit-cycle from which to borrow. Instead, here a variable delay line (VDL) has been inserted in series with the latch strobe signal to extend analog settling time in the first half of every bit-cycle, including the first, "pre-borrowing" time from that bit-cycle's own latch phase. The beginning of every bit period is synchronous with the sampling clock, and the latch strobing is determined by the setting of the VDL, which is tuned externally to see tradeoffs between extended settling time and ADC performance.

C. Technology Considerations

The SAR architecture's digital complexity directly benefits from the reduced feature sizes. Even though this ADC uses a fully static CMOS logic style, it still consumes less power than the highly customized logic, including dynamic registers, used in [18]. Care was taken throughout the digital logic to provide the maximum robustness in presence of delay variations.

The two analog blocks are well suited for integration in 65-nm CMOS with the following design considerations. For the same absolute device size, transistor matching improves in successive technology generations, allowing smaller total device area and capacitance in the comparators [28]; however, the matching is not improved for minimum size devices. Also, due to the reduced power supplies and decreased $g_m r_o$ of the short channel devices, it is difficult to get high gain in a single analog stage. The preamplifiers and latch use non-minimum length transistors to improve both the matching and output impedance. While this does increase device capacitance for the same g_m, there is minimal power impact because wiring parasitics dominate the total capacitance in the comparator.

The capacitor array is entirely passive, and its switching speed is improved with the shorter gate lengths. Because no analog-specific processing steps (e.g., a thin oxide for high density MiM capacitors) were used in fabrication the capacitors are formed using interdigitated metal comb capacitors. The capacitance is determined by fringing between adjacent metal lines, structures that have been shown to achieve similar densities to MiM capacitors with matching limits at greater than the 7-bit level [29]. The capacitance size is chosen according to the matching requirements discussed in Section III-A. The input voltage is constrained to between 0 and 400mV to allow sampling with a single standard-V_T NFET transistor without exceeding the process voltage limit of 1.2 V.

IV. Measurements

The ADC has been fabricated in a 65-nm CMOS technology; a die photograph is shown in Fig. 12. With a 91-kHz input sampled at 500 MS/s, the INL and DNL are −0.16/0.15 and −0.20/0.26 LSBs, respectively (Fig. 13). The split capacitor array suffers no linearity degradation as compared to a separate on-chip test channel with the conventional array. The split array uses 31% less power from the 400 mV reference voltage supply; the difference in energy savings from the theory presented above is due to the increased bottom-plate routing.

The delay line was tested using an on-chip delay detection circuit and varying the input differential voltage. Due to an un-

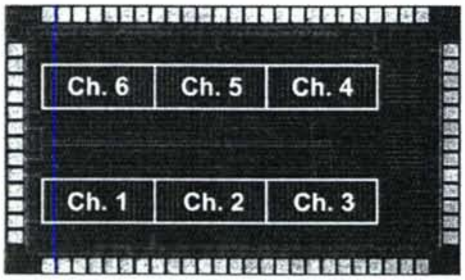


Fig. 12. Photograph of 1.9 × 1.4 mm die.

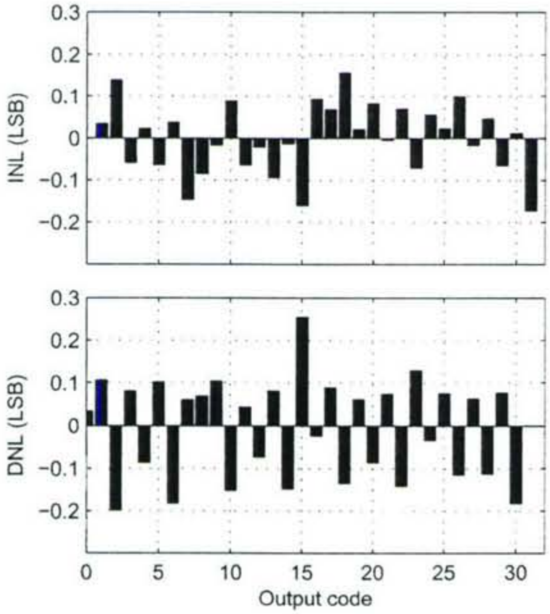


Fig. 13. Static linearity of ADC versus output code.

derestimation of parasitics in the delay line, only the first two delay steps out of 16 provided sufficient time for latch regeneration, and these extended the period available to the preamplifiers by about 10%. At 250 MS/s, a 0.5–1 dB improvement in SNDR was achieved by properly tuning the delay.

The dynamic performance of the ADC is shown in Fig. 14 with the input frequency swept from DC to beyond Nyquist. The signal-to-noise-plus-distortion ratio (SNDR) does not drop by 3 dB until past the Nyquist frequency. A fast Fourier transform (FFT) of a 239.04-MHz input is shown in Fig. 15. Spurs (a)–(d) result from gain errors and skew between channels, and spurs (e)–(f) are due to offset mismatch. All of these spurs are below −39 dBFS, and their combined power is still less than the total noise power (excluding the spurs) at this near-Nyquist input. The gain mismatch between channels is 0.9%. The individual channels have an effective number of bits (ENOB) between 4.65 and 4.75 with low-frequency inputs, dropping by 0.4 bits at Nyquist.

The ADC consumes 2.86 mW and 3.06 mW, respectively, from 1.2-V analog and digital supplies at the maximum sampling frequency. The ADC was also tested at lower sampling frequencies. At 250 MS/s, the ADC consumes a total of 1.58 mW from a 1 V digital and 0.8 V analog supply, while still maintaining Nyquist performance. A summary of the ADC is listed in Table I.

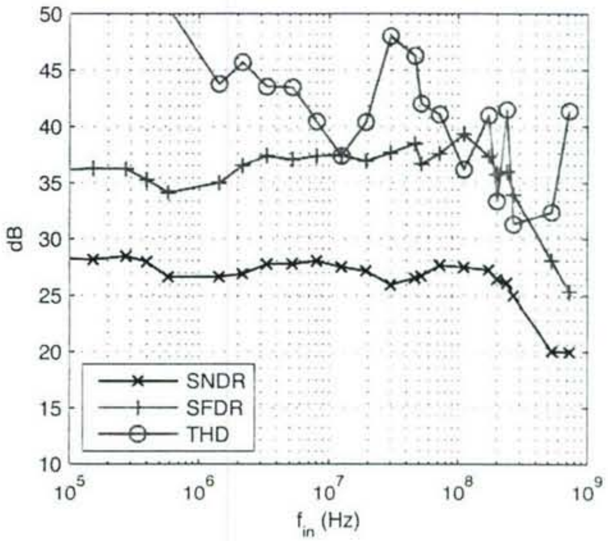


Fig. 14. Dynamic performance versus input frequency.

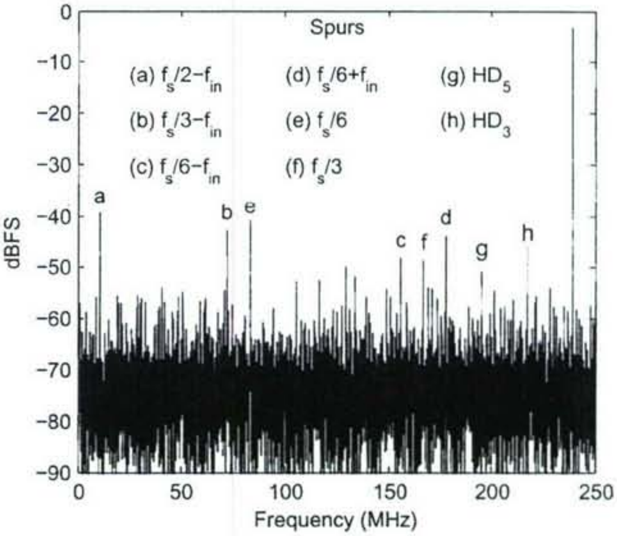


Fig. 15. FFT of 239.04-MHz sine wave sampled at 500 MS/s; dominant spurs are labeled.

TABLE I
SUMMARY OF PERFORMANCE

Technology	65-nm CMOS 1P6M
Supply Voltage	1.2 V
Sampling Rate	500 MS/s
Resolution	5 bit
Input Range	800 mV_{pp} Differential
SNDR (f_{in}=3.3 MHz)	27.8 dB
SNDR (f_{in}=239 MHz)	26.1 dB
SFDR (f_{in}=239 MHz)	36.0 dB
THD (f_{in}=239 MHz)	-41.5 dB
DNL (channel)	0.26 LSB
INL (channel)	0.16 LSB
Analog Power	2.86 mW
Digital Power	3.06 mW
Total Power	5.93 mW
Active Area	0.65 mm × 1.4 mm

TABLE II
COMPARISON OF STATE-OF-THE-ART ADCS

Work	Archite-cture	Feature Size	Power (mW)	f_S (MHz)	Res-olution (bits)	f_{in} (MHz)	ENOB	FOM (pJ/conv. step)
[17]	SAR	90 nm	10	600	6	300	5.1	0.5
[19]	SAR	0.13 μm	5.3	600	6	300	5.02	0.27
[30]	Subranging	0.13 μm	21	125	8	62.5	7.5	0.96
[13]	Pipelined	0.18 μm	30	200	8	99	7.68	0.74
[31]	Subranging	90 nm	55	1000	6	500	5.3	1.37
[32]	Flash	90 nm	2.5	1250	4	625	3.66	0.16
This work	SAR	65 nm	5.9	500	5	239	4.04	0.75
	SAR	65 nm	1.8	250	5	120	4.10	0.44
	SAR	65 nm	0.9	125	5	60	3.95	0.51

V. COMPARISON AND DISCUSSION

To enable a comparison to other ADCs operating at different speeds and resolutions, the figure of merit

$$\mathrm{FOM} = \frac{P}{2^{\mathrm{ENOB}} \cdot 2 \cdot f_{\mathrm{in}}} \tag{14}$$

is used [17], where P is the power consumption, and ENOB is measured for input frequency f_{in}, not to exceed Nyquist input. Table II compares state-of-the-art ADCs with sampling rates in excess of 100 MS/s and resolutions of 8 bits or less. From the results, this ADC has one of the best energy efficiencies of published work. In addition, as three out of the four best designs demonstrate, the time-interleaved SAR architecture can achieve very low power for these specifications. This work requires no linearity calibration or digital post-processing of the samples.

VI. CONCLUSION

An ADC targeted for UWB specifications has been presented. The time-interleaved SAR architecture provides superior energy efficiency to a flash converter because of its linear growth in complexity with the resolution. Two new techniques have enabled high-speed, low-power SAR operation. The split capacitor array offers both lower switching energy and improved settling speed as compared to the conventional array. Joint timing design of the analog and digital portions of the chip, as demonstrated with the adjustable latch strobing instant, can ease settling time requirements and use otherwise wasted idle time during bit-cycling. State-of-the-art energy efficiency and performance have been demonstrated with robust operation in deep-submicron CMOS.

ACKNOWLEDGMENT

The authors would like to thank Texas Instruments for fabricating the chip. They would also like to thank C. Mangelsdorf of Analog Devices for feedback on the latch-delay circuit and N. Verma from MIT for many discussions throughout the design process.

REFERENCES

[1] A. B. Batra *et al.*, Multi-Band OFDM Physical Layer Proposal for IEEE 802.15 Task Group 3a IEEE, P802.15-04/0493r0 [Online]. Available: http://grouper.ieee.org/groups/802/15/pub/04/15-04-0493-00-003a-multi-band-ofdm-cfp-document-update.zip

[2] R. F. Fisher *et al.*, DS-UWB Physical Layer Submission to 802.15 Task Group 3a IEEE, P802.15-04/0137r3 [Online]. Available: http://grouper.ieee.org/groups/802/15/pub/04/15-04-0137-04-003a-merger2-proposal-ds-uwb-update.doc

[3] P. P. Newaskar, R. Blazquez, and A. P. Chandrakasan, "A/D precision requirements for an ultra-wideband radio receiver," in *IEEE Workshop on Signal Processing Systems*, Oct. 2002, pp. 270–275.

[4] E. S. Saberinia *et al.*, "Analog to digital converter resolution of multiband OFDM and pulsed-OFDM ultra wideband systems," in *Proc. 1st Int. Symp. Control, Communications, and Signal Processing*, 2004, pp. 787–790.

[5] B. R. Razavi *et al.*, "Multiband UWB transceivers," in *Proc. IEEE Custom Integrated Circuits Conf.*, 2005, pp. 141–148.

[6] D. D. W. Wentzloff *et al.*, "System design considerations for ultra-wideband communication," *IEEE Commun. Mag.*, vol. 43, no. 8, pp. 114–121, Aug. 2005.

[7] F. S. Lee *et al.*, "A 3.1 to 10.6 GHz 100 Mb/s pulse-based ultra-wide-band radio receiver chipset," in *IEEE Int. Conf. Ultra-Wideband*, Sep. 2006, pp. 185–190.

[8] G. Geelen, "A 6 b 1.1 Gsample/s CMOS A/D converter," in *IEEE ISSCC Dig. Tech. Papers*, 2001, pp. 128–129.

[9] K. Sushihara and A. Matsuzawa, "A 7b 450 MSample/s 50 mW CMOS ADC in 0.3 mm^2," in *IEEE ISSCC Dig. Tech. Papers*, 2002, vol. 1, pp. 170–171.

[10] P. Scholtens and M. Vertregt, "A 6-bit 1.6-GSample/s flash ADC in 0.18-μm CMOS using averaging termination," *IEEE J. Solid-State Circuits*, vol. 37, no. 12, pp. 1599–1609, Dec. 2002.

[11] X. Jiang and M.-C. F. Chang, "A 1-GHz signal bandwidth 6-bit CMOS ADC with power-efficient averaging," *IEEE J. Solid-State Circuits*, vol. 40, no. 2, pp. 532–535, Feb. 2005.

[12] C. S. Sandner *et al.*, "A 6-bit 1.2-GS/s low-power flash-ADC in 0.13-μm digital CMOS," *IEEE J. Solid-State Circuits*, vol. 40, no. 7, pp. 1499–1505, Jul. 2005.

[13] H.-C. Kim, D.-K. Jeng, and W. Kim, "A 30 mW 8 b 200 MS/s pipelined CMOS ADC using a switched-opamp technique," in *IEEE ISSCC Dig. Tech. Papers*, 2005, pp. 284–285.

[14] S. G. Gupta *et al.*, "A 1 GS/s 11 b time-interleaved ADC in 0.13 μm CMOS," in *IEEE ISSCC Dig. Tech. Papers*, 2006, vol. 49, pp. 576–577.

[15] W. Black and D. Hodges, "Time interleaved converter arrays," *IEEE J. Solid-State Circuits*, vol. 15, no. 12, pp. 929–938, Dec. 1980.

[16] J. Yuan and C. Svensson, "A 10-bit 5-MS/s successive approximation ADC cell used in a 70-MS/s ADC array in 1.2-μm CMOS," *IEEE J. Solid-State Circuits*, vol. 29, no. 8, pp. 866–872, Aug. 1994.

[17] D. Draxelmayr, "A 6 b 600 MHz 10 mW ADC array in digital 90 nm CMOS," in *IEEE ISSCC Dig. Tech. Papers*, 2004, pp. 264–265.

[18] B. P. Ginsburg and A. P. Chandrakasan, "Dual scalable 500 MS/s, 5b time-interleaved SAR ADCs for UWB applications," in *Proc. IEEE Custom Integrated Circuits Conf.*, 2005, pp. 403–406.

[19] S.-W. M. Chen and R. W. Brodersen, "A 6-bit 600-MS/s 5.3-mW asynchronous ADC in 0.13-μm CMOS," *IEEE J. Solid-State Circuits*, vol. 41, no. 12, pp. 2669–2680, Dec. 2006.

[20] B. P. Ginsburg and A. P. Chandrakasan, "A 500 MS/s 5 b ADC in 65 nm CMOS," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2006, pp. 174–175.

[21] S. J. Jamal *et al.*, "A 10-bit 120-Msample/s time-interleaved analog-to-digital converter with digital background calibration," *IEEE J. Solid-State Circuits*, vol. 37, no. 12, pp. 1618–1627, Dec. 2002.

[22] J. McCreary and P. Gray, "All-MOS charge redistribution analog-to-digital conversion techniques," *IEEE J. Solid-State Circuits*, vol. SC-10, no. 12, pp. 371–379, Dec. 1975.

[23] B. P. Ginsburg and A. P. Chandrakasan, "An energy-efficient charge recycling approach for a SAR converter with capacitive DAC," in *Proc. IEEE Int. Symp. Circuits and Systems*, 2005, vol. 1, pp. 184–187.

[24] C.-H. Lin and K. Bult, "A 10-b, 500-MSample/s CMOS DAC in 0.6 mm," *IEEE J. Solid-State Circuits*, vol. 33, no. 12, pp. 1948–1958, Dec. 1998.

[25] A. Hastings, *The Art of Analog Layout*. Upper Saddle River, NJ: Prentice-Hall, 2001.

[26] B. P. Ginsburg and A. P. Chandrakasan, "Dual time-interleaved successive approximation register ADCs for an ultra-wideband receiver," *IEEE J. Solid-State Circuits*, vol. 42, no. 2, pp. 247–257, Feb. 2007.

[27] G. Promitzer, "12-bit low-power fully differential switched capacitor noncalibrating successive approximation ADC with 1 MS/s," *IEEE J. Solid-State Circuits*, vol. 36, no. 7, pp. 1138–1143, Jul. 2001.

[28] M. Pelgrom, H. Tuinhout, and M. Vertregt, "Transistor matching in analog CMOS applications," in *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, 1998, pp. 915–918.

[29] R. Aparicio and A. Hajimiri, "Capacity limits and matching properties of integrated capacitors," *IEEE J. Solid-State Circuits*, vol. 37, no. 3, pp. 384–393, Mar. 2002.

[30] J. M. Mulder *et al.*, "A 21 mW 8 b 125 MS/s ADC occupying 0.09 mm^2 in 0.13 μm CMOS," in *IEEE ISSCC Dig. Tech. Papers*, 2004, pp. 260–261.

[31] P. M. F. Figueiredo *et al.*, "A 90 nm CMOS 1.2 V 6b 1.2 GS/s two-step subranging ADC," in *IEEE ISSCC Dig. Tech. Papers*, 2006, pp. 568–569.

[32] G. Van der Plas, S. Decoutere, and S. Donnay, "A 0.16pJ/conversion-step 2.5mW 1.25GS/s 4b ADC in a 90nm digital CMOS process," in *IEEE ISSCC Dig. Tech. Papers*, 2006, vol. 49, pp. 566–567.

Brian P. Ginsburg (S'04) received the S.B. and M.Eng. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, in 2003. He is currently working toward the Ph.D. degree at MIT.

His research interests include analog-to-digital converters, optimization of mixed-signal circuits, and ultra-wideband radio circuits and systems.

Mr. Ginsburg was named a Siebel Scholar in 2003 and received the NDSEG Fellowship in 2004.

Anantha P. Chandrakasan (M'95–SM'01–F'04) received the B.S, M.S., and Ph.D. degrees in electrical engineering and computer sciences from the University of California, Berkeley, in 1989, 1990, and 1994, respectively.

Since September 1994, he has been with the Massachusetts Institute of Technology, Cambridge, where he is currently the Joseph F. and Nancy P. Keithley Professor of Electrical Engineering. His research interests include low-power digital integrated circuit design, wireless microsensors, ultra-wideband radios, and emerging technologies. He is a coauthor of *Low Power Digital CMOS Design* (Kluwer, 1995) and *Digital Integrated Circuits* (Pearson Prentice-Hall, 2003, 2nd edition). He is also a co-editor of *Low Power CMOS Design* (IEEE Press, 1998), *Design of High-Performance Microprocessor Circuits* (IEEE Press, 2000), and *Leakage in Nanometer CMOS Technologies* (Springer, 2005).

Dr. Chandrakasan has received several awards including the 1993 IEEE Communications Society's Best Tutorial Paper Award, the IEEE Electron Devices Society's 1997 Paul Rappaport Award for the Best Paper in an EDS publication during 1997, the 1999 Design Automation Conference Design Contest Award, and the 2004 DAC/ISSCC Student Design Contest Award. He has served as a technical program co-chair for the 1997 International Symposium on Low Power Electronics and Design (ISLPED), VLSI Design'98, and the 1998 IEEE Workshop on Signal Processing Systems. He was the Signal Processing Subcommittee Chair for ISSCC 1999–2001, the Program Vice-Chair for ISSCC 2002, the Program Chair for ISSCC 2003, and the Technology Directions Subcommittee Chair for ISSCC 2004–2006. He was an Associate Editor for the IEEE JOURNAL OF SOLID-STATE CIRCUITS from 1998 to 2001. He serves on the SSCS AdCom and is the meetings committee chair. He is the Technology Directions Chair for ISSCC 2007.